From Failure Taxonomy to Intervention: A Diagnostic Methodology for Industry-Scale AVLM in Video and Live-Streaming Platform Moderation

Shuchang Ye[1,2] **Jinqiang Yu**[1] **Zhujun Xiao**[1] **Yajing Kong**[1]
Yist Y. Lin[1] **Yang Ma**[1] **Jiaxi Liu**[1] **Xiaolei Xu**[1] **Zheng Yu**[1]

[1]TikTok [2]The University of Sydney

Abstract

Industry-scale video and live-streaming moderation imposes requirements that are difficult to satisfy with generic pretrained public models or external APIs, including adaptation to platform-specific data distributions, policy-specific objectives, and product-level safety constraints. As a result, platforms must undertake internal model development, naturally turning to shared public research for guidance. However, existing multimodal foundation-model studies primarily report architectures, training recipes, data scaling strategies, and benchmark results, but provide less systematic guidance on how failures should be localized and translated into targeted model-development interventions. Interventions are essential because deployment failures are rarely self-explanatory. Similar failures can originate from different causes. Without targeted interventions, improvement reduces to heuristic trial-and-error, where benchmark improvements are weakly attributable, and failures are difficult to trace to their underlying causes. To address this gap, we present a diagnostic methodology for industry-scale Audio-Visual-Language Models (AVLM) development. The methodology maps model failures into a taxonomy of observable failure signatures and links each class of failure to an intervention space. We instantiate this methodology through the development and alignment of the first AVLM foundation model trained on trillion-scale data for general content understanding and downstream moderation on video and live-streaming platforms. The resulting system is deployed in production across more than 100 regions and more than 100 languages, on a platform serving billions of users and supporting tens of billions of quarterly video uploads and live streaming.

1 Introduction

Modern video and live-streaming platforms host massive volumes of user-produced multimodal content, making automated moderation a core infrastructure requirement for maintaining community safety at scale [1, 12, 25]. The scale of this requirement is evident from industry transparency disclosures. Earlier reports emphasized the human labor burden of moderation, describing workforces of approximately $15{,}000$ reviewers [26]. Whereas more recent disclosures highlight the degree of operational automation, reporting roughly 112 million pieces of violating content removed within six months, with 93.8% actioned without human review [30]. In policy-sensitive moderation, externally hosted APIs and generic pretrained models provide limited control over the operational and behavioral factors that determine deployment quality, including latency, cost, failure modes, and policy alignment. This motivates the development of platform-controlled Audio-Visual-Language Models (AVLMs) that can be optimized for reliability, scalability, and alignment with evolving moderation policies.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

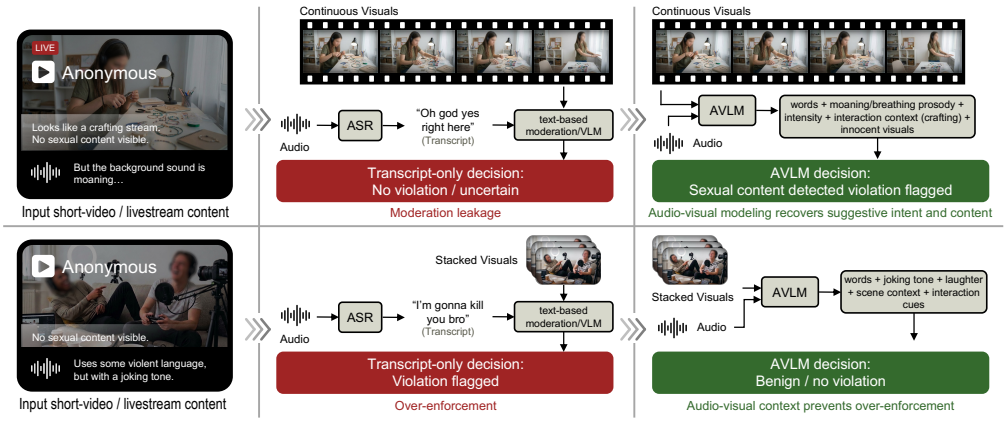

Figure 1: Audio evidence beyond transcription in video and live-streaming moderation. The upper example illustrates moderation leakage when policy-relevant acoustic evidence is lost during transcription. The lower example illustrates overkill when risky words are preserved in transcript form but the original audio-visual context indicates benign intent. (*Privacy is protected through generative face anonymization.*)

Existing AI-assisted moderation systems for video and live-streaming platforms commonly adopt Vision-Language Model (VLM)-based pipelines [34], where visual frames serve as the primary perceptual input for policy reasoning. In such pipelines, audio is often represented indirectly through ASR transcripts, which convert speech into textual evidence for the moderation model. This representation captures violations expressed through recognized words, but neglects acoustic and cross-modal evidence such as prosody, vocal intensity, non-speech sounds, environmental audio, and correspondences between sound and visual context. Figure 1 illustrates two failure modes: Transcript-based systems may under-enforce when policy-relevant evidence is present in the audio signal but absent from the transcript, or over-enforce when risky words are disambiguated by the original audio-visual context.

These limitations motivate a shift from transcript-based audio modeling to Audio-Visual-Language Models (AVLMs) that jointly integrate speech, non-speech acoustic cues, visual context, and contextual metadata for moderation. However, existing literature in multimodal foundation models primarily focuses on model-building procedures, emphasizing architecture design, modality alignment, data scaling, staged training, and benchmark performance. Representative works have established visual instruction tuning for VLM alignment [20], scalable transfer across image and video scenarios [15], multimodal instruction tuning over video and audio [5], and general-purpose omni-modal modeling [36]. A central deployment question remains rarely discussed: *How can model failures be diagnosed in ways that identify appropriate interventions?*

This question is important in industry internal model development because downstream policy errors often fail to reveal their underlying cause. A single moderation failure could stem from a variety of distinct technical issues. Without a systematic diagnostic framework, performance improvement can be reduced to heuristic trial-and-error, where benchmark improvements are weakly attributable, and failures are difficult to trace to their underlying causes.

To address this gap, we introduce a systematic diagnostic methodology for industry-scale AVLM development, in which observed failures are categorized into interpretable failure taxonomies and translated into actionable development interventions. We instantiate this methodology across the development and alignment lifecycle of an AVLM foundation model for a large-scale video and live-streaming platform. The resulting system supports over 100 regions and is designed for noisy, ambiguous, and highly diverse content drawn from global platform traffic.

2 AVLM Development for Industry-Scale Video and Live-Streaming Platforms

In this section, we present an overview of the development pipeline of AVLM, as shown in Figure 2.

2.1 Supervision Regimes for Platform Audio

Training data is categorized into three supervision regimes that progressively shift from scalable audio acquisition to task-oriented multimodal reasoning, as shown in Table 1.

Weakly supervised audio-text data. The first regime contains approximately **1 million hours** of pseudo-labeled audio-text pairs from online ASR sources. It provides broad multilingual speech coverage for initial audio capability acquisition. We apply automatic quality control, including agreement-based filtering between audio language identification and ASR outputs, and entropy-based filtering to remove repetitive or low-information transcripts.

Supervised audio-only data. The second regime contains approximately **200,000 hours** of supervised audio data from in-house video and live scenarios. It moves training beyond transcript reconstruction toward broader auditory understanding, covering ASR, AST, diarization-ASR, speech summarization, intent understanding, harmful-content identification, and audio event summarization. Compared with the weakly supervised corpus, this regime is smaller but provides denser supervision for speaker behavior, paralinguistic cues, and mixed speech/non-speech acoustic patterns.

Supervised audio-visual instruction data. The third regime introduces audio-instruction following and cross-modal collaboration, aiming to integrate audio-visual reasoning while preserving established visual-textual performance. Transcript pairs broaden lexical exposure, whereas event and intent targets make acoustic properties useful beyond transcription. Audio-visual targets additionally require interpreting whether the modalities reinforce or disambiguate each other.

Table 1: Training supervision regimes used in the AVLM development pipeline.

Regime	Primary tasks
T0: weak audio-text data	ASR-style pseudo-labeling
T1: supervised audio data	ASR, audio event perception, diarization, AST
T2: supervised audio-visual data	AV grounding, VQA/AVQA, instruction following

2.2 Architecture

The AVLM couples an in-house audio encoder with a pretrained language or vision-language backbone via a lightweight projector. Raw waveform segments are encoded into audio representations, mapped into the backbone embedding space, and jointly processed with text tokens and, visual tokens from sampled frames. This formulation enables speech content, non-speech acoustic cues, visual context, and cross-modal interactions to be modeled within a shared multimodal language-model decoder. The encoder must retain relevant acoustic cues, the projector must make them accessible, and the decoder must use them. Reliable lexical recovery alone does not test whether this chain supports other policy-relevant evidence.

2.3 Three-stage AVLM Training

Stage I. Weakly supervised audio text alignment. The first stage establishes the audio language alignment using large scale weakly supervised audio text data. Its objective is to map audio encoder representations into the pretrained language backbone space and acquire broad audio text correspondence from high coverage pseudo supervision.

Stage II. Audio perception and reasoning adaptation. The second stage adapts the audio encoder and projection modules with supervised audio data. This stage strengthens speech understanding, non-speech perception, and audio reasoning beyond transcript reconstruction. It also improves

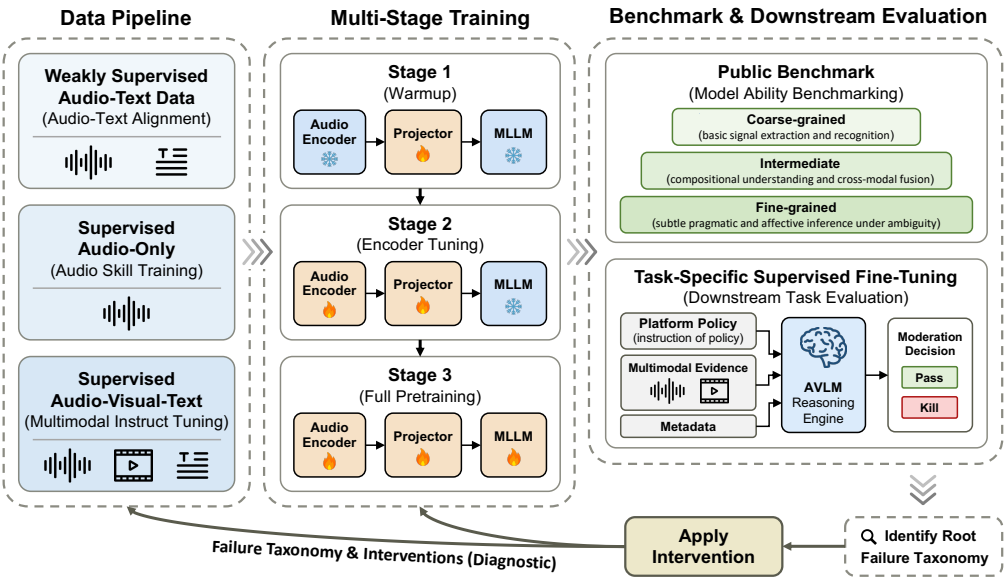

Figure 2: Overview of the proposed AVLM development and diagnostic framework. The framework couples a staged data pipeline, multi-stage training, and benchmark&downstream evaluation for industrial AVLM development. It also exposes a diagnostic feedback loop, in which observed evaluation failures are abstracted into root failure taxonomies and mapped to interventions.

sensitivity to paralinguistic cues, speaker behavior, and environmental sounds, thereby reducing over-reliance on ASR-style speech supervision.

Stage III. Joint multimodal capability consolidation. The third stage trains the full model using mixed supervision spanning audio, vision, text, and audio-visual instructions. This stage aims to preserve the vision-language capabilities inherited from the VLM backbone while further improving audio understanding and audio-visual reasoning. Full-model updates change the decoder interface against which audio representations were learned. Each transition therefore requires checking both newly acquired capability and regression on earlier tasks.

Foundation-model pretraining uses one epoch, a learning rate of 10^{-5}, and global batch size 1,024. Moderation fine-tuning uses two epochs, a learning rate of 5×10^{-6}, per-device batch size 2, and eight gradient-accumulation steps.

3 Experiment Setup

Our experiments evaluate an in-house AVLM designed for audio perception, audio-visual reasoning, and platform-oriented multimodal understanding. We evaluate the model in two dimensions: general multimodal understanding and downstream moderation performance.

3.1 Evaluation Datasets

For downstream moderation evaluation, we use anonymized in-house video and live-streaming benchmarks covering two primary policy families, denoted as *Anon.S* representing sexual-related violation and *Anon.V* representing violence-related violation. These benchmarks evaluate policy-aligned audio-visual AI moderation performance under large-scale complex-context real-world scenarios.

For public audio-visual evaluation, we use AVHBench [29] and OmniBench [19]. These benchmarks evaluate zero-shot audio-visual understanding, audio-visual matching, omni-modal reasoning, and instruction-following behavior.

For audio-only evaluation, we use FLEURS [6], MMAU [27], and MMSU [31]. FLEURS evaluates multilingual speech recognition across 18 languages. MMAU and MMSU evaluate broader audio perception, reasoning, and knowledge-oriented understanding.

For general multimodal evaluation, we use benchmarks covering STEM reasoning, general VQA, document and chart understanding, OCR, counting, multi-image reasoning, spatial understanding, and video reasoning. These include MMMU [37], RealWorldQA [35], MMStar [4], GQA [13], SEED-Bench [16], POPE [18], ScienceQA [22], MME [9], AI2D [14], OCRBench [21], CountBench [2], DocVQA [24], ChartQA [23], TextVQA [28], InfoVQA [3], BLINK [11], MUIRBENCH [32], ERQA [8], EmbSpatialBench [7], MVBench [17], VideoMME [10], and LVBench [33].

3.2 Metrics

For downstream moderation benchmarks, we report PR-AUC because the evaluated policy behaviors are class-imbalanced. For FLEURS, we report WER/CER, where lower values indicate better multilingual speech recognition. For MMAU, MMSU, AVHBench, OmniBench, and most general multimodal benchmarks, we report the benchmark-defined accuracy or score. For document-oriented benchmarks, we use the benchmark-standard metrics, including ANLS and exact match where applicable.

4 Failure Taxonomy and Interventions

This section presents representative failure taxonomies observed during AVLM development and maps each failure type to targeted interventions within the development lifecycle. These categories can overlap. Auditing retained support before testing further exposure separates missing examples from under-training. Re-evaluation after instruction tuning then distinguishes new regression from earlier supervision bottlenecks.

4.1 Selection-Induced Support Erosion

In the Stage I pretraining data pipeline, large weakly supervised corpora are typically collected from platform traffic to provide broad coverage of real-world, user-produced content. However, this raw data is systematically reshaped by media availability, validity checks, audio-visual alignment constraints, and quality-control filters. Because these selection mechanisms are not distribution-neutral, they preferentially retain clean, high-resource, and easily aligned examples. Consequently, they actively suppress the noisy, long-tail cases that are critical for learning nuanced moderation boundaries.

Let $p_{\text{traffic}}(x)$ denote the underlying distribution of raw platform content. The data pipeline subjects each sample x to a sequence of filters, which we can represent as an aggregate retention probability $S_{\text{retain}}(x)$. The final, post-filter training distribution $q(x)$ is therefore:

$$q(x) = \frac{p_{\text{traffic}}(x) S_{\text{retain}}(x)}{\mathbb{E}_{x \sim p_{\text{traffic}}}[S_{\text{retain}}(x)]}$$

When $S_{\text{retain}}(x)$ is lower for minority languages, poor audio conditions, or complex multimodal edge cases, raw corpus size can overstate the coverage available for learning. Therefore, diagnostics and interventions must target the post-filter training distribution $q(x)$, rather than $p_{\text{traffic}}(x)$. Table 2 summarizes the diagnostic signals of this erosion and their corresponding controls. We compare pre- and post-filter distributions by region, language, audio quality, supervision source, and policy tails. Reweighting can increase exposure to retained cases, but cannot recover a case type with zero retained support. That requires revising collection or filtering.

***Failure Taxonomy* 1.** Data curation pipelines act as non-neutral filters, transforming broad raw traffic into a narrow training distribution. This strips away the critical long-tail and noisy examples required for robust moderation capability.
***Intervention*.** Treat the data pipeline as a biased sampling mechanism. Reconstruct the necessary distribution through region-aware sampling, content-driven upweighting, and calibrating quality filters to retain policy-relevant edge cases.

Table 2: Diagnostic signals for selection-induced support erosion. Controls must target the post-selection training distribution rather than the raw collection corpus.

Diagnostic Signal	**Underlying Mechanism**	**Corrective Control**
Highly skewed region and language histograms	Filtering over-represents high-resource demographics, degrading minority-language and region-specific acoustic performance.	Post-filter region and language balancing.
Raw traffic is dominated by benign "head" content	Abundant supervision on frequent, low-risk patterns crowds out exposure to rare, high-risk policy violations.	Content-aware sampling and upweighting of tail events.
Disproportionate removal of weakly aligned media	Strict quality control systematically deletes the noisy, ambiguous samples that models need for robust real-world deployment.	Alignment-aware filter calibration and post-filter auditing.

4.2 Regime-Limited Audio-Language Alignment

Weak audio-text pretraining in Stage I serves as the first diagnostic point in AVLM development because it determines whether acoustic representations can be reliably interfaced with the language backbone. The objective of this stage is not limited to transcription accuracy. It is to establish an audio-language mapping that later stages can inherit for reasoning, instruction following, and multimodal grounding. When this pretraining is not converged, failures observed in later supervised stages may originate from unstable cross-modal representations rather than from insufficient reasoning supervision.

Figure 3 illustrates this transition through the scaling behavior of weak audio-text pretraining. Continued held-out improvement supports extending weak pretraining. A stable plateau motivates testing higher-information supervision, but does not by itself identify a unique bottleneck. Strong transcription with persistently weak audio question answering (AQA) after this plateau distinguishes the next diagnosis from insufficient Stage-I exposure.

***Failure Taxonomy* 2.** The audio-language mapping remains unreliable when weak pretraining has not reached the exposure regime required for stable cross-modal alignment.
***Intervention*.** Use scaling curves as transition tests: scale weak pretraining while alignment error decreases, then replace additional exposure with higher-information reasoning and multimodal supervision.

4.3 Proxy-Sufficient Representation Collapse

After Stage I weak audio-text pretraining is saturated, the next diagnostic question is whether the learned audio representation contains information that is usable by the decoder beyond lexical recovery. Stage I audio-to-text pretraining rewards representations that preserve lexical content, but may underweight semantic intent, paralinguistic cues, non-speech events, background context, and audio-visual consistency. These non-speech signals are often necessary for Stage II audio reasoning and downstream moderation.

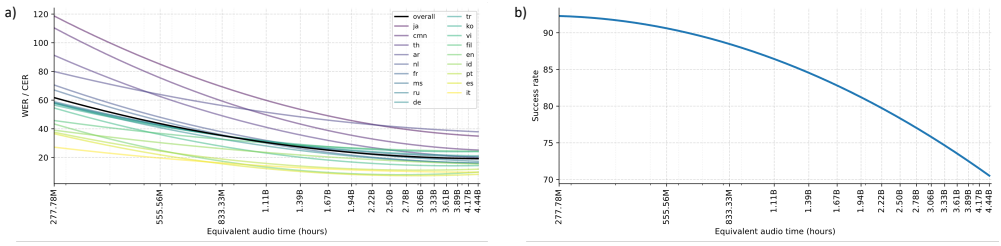

Figure 3: Effect of scaling equivalent audio training time. a) FLEURS WER/CER across common languages and overall. b) MMSU question-answering success rate.

This mismatch can make the audio pathway overfit to speech recognition. The model is optimized for recognizing lexical content, as reflected by lower WER or CER, but the resulting audio features may not preserve sufficient audio information for downstream reasoning, instruction following, or cross-modal grounding. As a result, ASR performance can improve while AQA accuracy degrades.

Cross-Modal Continuation [1] (CMC) addresses this representation-level mismatch before Stage III instruction tuning by requiring the decoder to continue language from the audio context. The objective does not directly optimize downstream policy decisions. Instead, it aligns projected audio representations with the language-generation interface of the backbone, with the aim of helping the audio pathway use the instruction-following and reasoning capabilities of the backbone. Table 3 shows that CMC improves AQA performance while preserving or strengthening transcription quality. CMC uses timestamp-aligned pairs split at a valid word boundary, with an audio prefix as input and transcript suffix as target. Low-confidence alignments and too-short examples are removed. CMC comprises 20% of the audio-data mixture. Joint ASR and AQA improvements show no inherent conflict, but do not isolate representation changes from data selection or continuation supervision.

Table 3: Effect of cross-modal continuation (CMC) on transcription and AQA-style capability.

Model	**LibriSpeech-Other** ↓	**FLEURS overall** ↓	**MMAU** ↑	**MMSU** ↑
Without CMC	13.40	29.26	34.2	18.98
With CMC	10.41	17.98	52.5	40.98

***Failure Taxonomy* 3.** A representation can become sufficient for a dominant proxy objective while under-encoding information needed for reasoning, instruction following, and decision use.
***Intervention*.** Add cross-modal generative supervision that trains the modality representation to support decoder continuation beyond proxy recovery.

4.4 Instruction Tuning Can Trade Perception for Reasoning

Stage III task-oriented training emphasizes instruction-following and multimodal reasoning. A typical failure mode is that reasoning and instruction-following abilities improve while perception is weakened. Because the model is optimized by response-level supervision, where correct answers may be obtained from higher-level semantic cues, visual context, metadata, or dataset priors without requiring full retention of speech-perception fidelity.

Table 4 shows that instruction tuning improves AQA reasoning but can degrade ASR performance. CMC partially mitigates this degradation, consistent with aligning the intermediate audio representation with the language-model decoding space before Task-oriented training. Since CMC trains the model to perform language continuation from audio context, it encourages the audio pathway to preserve language-predictive lexical, semantic, and acoustic evidence. This may provide a stronger initialization for instruction tuning and reduce the risk that Stage III optimization relies on shortcuts that bypass detailed audio perception. FLEURS error remains 33.87 with CMC after Stage III, versus 29.26 in the Stage-II baseline and 17.98 at the pre-instruction CMC checkpoint in Table 3. Thus, mitigation is partial, and improved reasoning does not imply retention of the strongest earlier speech capability.

Table 4: Instruction tuning improves AQA reasoning while weakening speech perception.

Checkpoint	**LibriSpeech-Other** ↓	**FLEURS overall** ↓	**MMAU** ↑	**MMSU** ↑
Stage II	13.40	29.26	34.2	18.98
Stage III without CMC	13.66	34.90	71.0	51.06
Stage III with CMC	12.89	33.87	72.4	51.34

[1] Cross-modal continuation refers to generating or extending content in one modality conditioned on context from another modality. In our case, the model generates text conditioned on audio.

***Failure Taxonomy* 4.** Instruction tuning can trade perception for instruction following and reasoning. High-quality multimodal instruction data improves instruction-following and reasoning while weakening perception.
***Intervention*.** Apply cross-modal continuation before Stage III to strengthen decoder-aligned audio conditioning.

4.5 The Perception Assumption: Does Better ASR Guarantee Better Moderation?

Across Stage II audio adaptation and Stage III multimodal consolidation, an intuitive hypothesis is that improving fundamental audio perception, essentially making the model a better transcriber, will result in better downstream AI moderation. This assumption is inherited from legacy safety pipelines, where audio moderation is typically handled by cascading an ASR module to convert audio signals into textual transcripts before applying a VLM or LLM.

However, downstream AI moderation relies on reasoning about semantic intent, paralinguistic evidence, acoustic context, audio-visual grounding, and policy-specific decision boundaries. Consequently, a model can significantly improve its generic perception capabilities while actively degrading its downstream moderation performance, as illustrated in Table 5. The architecture, initialization, optimization, and total budget are fixed while ASR, non-ASR audio, and AQA shares change. The result therefore identifies a recipe-level trade-off. WER/CER weights lexical errors without measuring their policy consequences, so it cannot alone determine the best moderation recipe.

Table 5: Recipe-level trade-off under fixed training budget. The perception-heavy recipe emphasizes lexical recovery. The reasoning-balanced recipe upweights non-ASR audio tasks and AQA. LS denotes LibriSpeech.

Recipe	**ASR ↓**			**AQA ↑**		**Moderation ↑**	
	LS-Clean	**LS-Other**	**FLEURS**	**MMAU**	**MMSU**	**Anon.S**	**Anon.V**
Perception-heavy	7.21	12.85	20.29	43.0	40.0	0.459	0.517
Reasoning-balanced	7.91	12.89	29.90	72.5	51.34	0.565	0.580

***Failure Taxonomy* 5.** A model can improve generic perception ability while degrading downstream AI moderation, although this comparison does not establish competition within the learned representation.
***Intervention*.** Frame model selection as a multi-objective optimization problem. Target the Pareto frontier between generic perception and complex reasoning, requiring minimum capability scores before selecting for downstream moderation performance.

5 Evaluation

We systematically evaluate our AVLM on public benchmarks and anonymized downstream moderation tasks, covering audio-visual understanding, audio capability, and visual capability. To contextualize performance under comparable model scale, we compare against Qwen2.5-Omni-3B and Gemma-4-E2B-it, two recent open-source multimodal baselines with similar parameter sizes.

5.1 Overall Multimodal Evaluation

Table 6 reports public audio-visual, audio, and visual benchmark results against Qwen2.5-Omni-3B and Gemma-4-E2B-it. Under a comparable model-scale setting, our AVLM demonstrates strong overall competitiveness, with particularly clear advantages in multilingual speech recognition and visual-language understanding. Despite using a compact 2B-scale architecture, our model achieves the lowest FLEURS WER/CER among the three models, showing that strong speech recognition is achievable at this model scale. The model also shows strong visual capability across general VQA, text/chart/document understanding, multi-image reasoning, and spatial reasoning. Establishing retention of inherited capability requires a comparison with the unadapted backbone. While Qwen2.5-Omni-3B remains stronger on several audio-reasoning, audio-visual, and video-centric benchmarks,

Table 6: Overall multimodal evaluation on public benchmarks. Benchmarks are grouped by main modality category and visual subcategory. Bold indicates the best result among the three models.

Category	Subcategory	Benchmark	Metric	Our AVLM-2B	Qwen2.5-Omni-3B	Gemma-4-E2B-it
Audio-Visual	AV Understanding	AVHBench	Accuracy	61.60	**71.00**	50.72
		OmniBench	Accuracy	43.86	**43.96**	21.71
Audio	Speech Recognition	FLEURS	WER/CER ↓	**13.45**	41.15	16.03
	Audio Understanding	MMAU	Accuracy	68.30	**72.80**	55.90
	Speech Understanding	MMSU	Accuracy	51.90	**61.40**	52.80
Visual	STEM	MMMU	Accuracy	42.67	**44.56**	34.78
	General VQA	RealWorldQA	Accuracy	**62.75**	58.69	43.92
		MMStar	Avg. accuracy	**57.10**	53.82	28.85
		GQA	Accuracy	**58.40**	26.46	39.51
		SEEDBench	Accuracy	75.10	**76.17**	64.71
		POPE	Accuracy / F1	**90.29 / 90.20**	88.23 / 87.93	84.01 / 84.07
		ScienceQA	Accuracy	**85.03**	81.66	51.21
		MME	Score (perc. + cog.)	1562 + 493	**1627 + 560**	938 + 227
	Text/Chart/Doc	AI2D	Accuracy	74.00	**76.98**	24.84
		OCRBench	Score	75.90	**78.20**	72.80
		CountBench	Accuracy	**86.97**	74.13	52.14
		DocVQA	ANLS	**91.51**	38.27	73.49
		ChartQA	Relaxed accuracy	**77.72**	19.84	51.72
		TextVQA	VQA accuracy	**75.52**	20.17	61.19
		InfoVQA	ANLS	**68.18**	36.33	39.12
	Multi-Image	BLINK	Accuracy	**46.93**	36.39	43.93
		MUIRBench	Accuracy	**46.88**	35.15	21.58
	Spatial	ERQA	Accuracy	**39.75**	36.75	35.50
		EmbSpatialBench	Accuracy	**62.72**	58.16	48.32
	Video	MVBench	Accuracy	59.00	**65.50**	25.24
		VideoMME w/o subtitles	Accuracy	54.81	**62.37**	42.00
		LVBench	Accuracy	38.22	**46.48**	24.92

and Gemma-4-E2B-it shows competitive performance on selected speech-understanding tasks, the overall results show that our AVLM provides a favorable balance between model size, audio capability, and multimodal competence. The transcription advantage alongside weaker AQA and video scores makes interpretation of temporal multimodal evidence a separate development target. Recognition accuracy alone does not establish the reasoning needed for moderation.

5.2 Downstream Modality Ablation

To evaluate the impact of audio modeling on downstream moderation, we conduct a modality ablation study on anonymized platform data collected across over 100 regions. We compare a transcript-based baseline (`ASR+T`), which processes ASR-transcribed speech alongside contextual metadata, against audio modeling (`A+T`) that replaces the transcript with parsed direct real audio signals while retaining the same metadata. We evaluate `A+T` under both frozen-backbone and fully adapted configurations. Anon.S and Anon.V contain approximately 45,500 and 117,403 samples, with 12% and 8% positives, from three months of traffic across over 100 languages. Each sample receives at least three independent annotations, with up to six for ambiguous cases. High-confidence positive labels must lead the next-highest label by at least two votes.

As shown in Table 7, audio modeling yields higher observed PR-AUC compared to the transcript baseline, with full adaptation yielding the strongest performance. These gains are particularly pronounced for `Anon.S`, where policy-relevant evidence often depends on paralinguistic and non-lexical cues. Frozen audio improves Anon.S/Anon.V by 0.041/0.003, while full adaptation adds 0.066/0.018. Thus, input substitution alone does not account for the full improvement.

We also observe improvements on `Anon.V`, which is consistent with audio providing complementary evidence, such as tone, acoustic intensity, and background noise, beyond textualized speech. These findings are compatible with cascaded transcript-based moderation being vulnerable to both information loss from discarding critical acoustic evidence and context collapse from evaluating risky text without the necessary paralinguistic framing. The ablation does not isolate acoustic mechanisms from transcription errors or policy adaptation. Bootstrap intervals quantify evaluation-set uncertainty, with paired intervals needed for between-setting differences.

Table 7: Transcript-based and audio-based moderation on the held-out policy sets. Entries show PR-AUC and aggregate 95% bootstrap confidence intervals.

Setting	Anon.S PR-AUC ↑	Anon.V PR-AUC ↑
ASR+T	0.447 [0.402, 0.495]	0.559 [0.520, 0.599]
A+T (freeze VL)	0.488 [0.442, 0.534]	0.562 [0.525, 0.602]
A+T (full)	0.554 [0.509, 0.599]	0.580 [0.542, 0.619]

5.3 Multilingual Speech Recognition

Table 8 reports per-language multilingual speech recognition results on FLEURS. Our AVLM obtains the best average WER/CER among the three models, with an average error rate of 13.45 compared with 41.15 for Qwen2.5-Omni-3B and 16.03 for Gemma-4-E2B-it. Across the 18 evaluated languages, our AVLM achieves the best result on 13 languages. Qwen2.5-Omni-3B performs best on English and Chinese, while Gemma-4-E2B-it performs best on Arabic, Thai, and Turkish. Overall, the results show that our audio training pipeline provides strong multilingual speech recognition performance while maintaining competitive audio-visual and visual-language capability. Arabic error of 45.50 versus 22.05 for Gemma-4-E2B-it nevertheless motivates language-specific checks of retained training support and moderation performance. The aggregate advantage does not establish uniform reliability across languages.

Table 8: FLEURS WER/CER (↓) across 18 languages. AVLM, Qwen, and Gemma denote Our AVLM-2B, Qwen2.5-Omni-3B, and Gemma-4-E2B-it. Bold indicates the best result per language.

Lang.	AVLM	Qwen	Gemma	Lang.	AVLM	Qwen	Gemma	Lang.	AVLM	Qwen	Gemma
en	6.50	**4.18**	9.14	ko	**15.97**	56.07	22.96	ms	**16.98**	36.02	17.49
ja	**18.71**	50.33	24.70	id	**11.79**	27.46	13.04	fil	**18.30**	67.14	21.87
de	**8.44**	28.90	13.44	th	10.15	54.97	**9.63**	ru	**8.09**	64.17	14.53
es	**6.18**	33.71	9.53	fr	**9.44**	22.58	14.88	tr	20.35	75.96	**16.52**
ar	45.50	70.30	**22.05**	vi	**11.69**	37.94	17.80	zh	11.25	**5.33**	21.50
pt	**5.47**	36.38	10.95	it	**4.66**	14.79	9.82	nl	**12.70**	54.54	18.76
Average	**13.45**	41.15	16.03								

6 Conclusion

This work presents a failure taxonomy for industry-scale AVLM development pipelines, categorizing recurring model failures and associating each failure class with targeted interventions. By integrating these interventions throughout the AVLM construction and alignment process, the resulting model demonstrates improved general audio-visual capability, stronger adaptation to platform-specific content, and consistent gains on downstream AI moderation tasks.

The foundation model developed in this work has since been adopted for production moderation on a global video and live-streaming platform. Sharing the development experience behind this deployment provides a practical account of the decisions and challenges involved in bringing an AVLM foundation model into industrial use.

We hope this study encourages the community to recognize model behavior intervention as an important research problem and to pursue further work in this direction.

Limitations. Due to platform privacy, safety, and policy-governance constraints, some production datasets, policy definitions, and sensitive moderation examples cannot be publicly released, and several downstream evaluations are therefore reported using anonymized policy categories.

Acknowledgements and Disclosures

This work was conducted at and supported by TikTok. All authors were employed by TikTok during the conduct of this research. The authors declare no additional competing interests.

References

[1] Nouar AlDahoul, Myles Joshua Toledo Tan, Harishwar Reddy Kasireddy, and Yasir Zaki. Advancing content moderation: Evaluating large language models for detecting sensitive content across text, images, and videos. *arXiv preprint arXiv:2411.17123*, 2024.

[2] Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. Countgd: Multi-modal open-world counting. *Advances in Neural Information Processing Systems*, 37:48810–48837, 2024.

[3] Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. Screenai: A vision-language model for ui and infographics understanding. *arXiv preprint arXiv:2402.04615*, 2024.

[4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

[5] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[6] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

[7] Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355, 2024.

[8] FlagEval. ERQA. `https://huggingface.co/datasets/FlagEval/ERQA`, 2025. Hugging Face dataset repository adapted from embodiedreasoning/ERQA. Accessed: 2026-05-03.

[9] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[10] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025.

[11] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[12] Tao Huang. Content moderation by llm: from accuracy to legitimacy. *Artificial Intelligence Review*, 58(10):320, 2025.

[13] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[14] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016.

[15] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[16] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[17] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[18] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305, 2023.

[19] Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024.

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

[21] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[22] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35: 2507–2521, 2022.

[23] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.

[24] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[25] Konstantina Palla, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Andreas Damianou, Henrik Lindström, Dan Taber, and Mounia Lalmas. Policy-as-prompt: Rethinking content moderation in the age of large language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 840–854, 2025.

[26] Reuters Staff. Facebook's flood of languages leaves it struggling to monitor content. *Reuters*, May 2019. URL `https://www.reuters.com/article/world/facebooks-flood-of-languages-leaves-it-struggling-to-monitor-content-idUSKCN1RZ0DL/`. Accessed: 2026-03-17.

[27] Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

[28] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[29] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024.

[30] TikTok. Our sixth transparency report on content moderation in europe. https://newsroom.tiktok.com/digital-services-act-our-sixth-transparency-report-on-content-moderation-in-europe?lang=en-150, February 2026. Reports around 112 million violating pieces of content removed in July–December 2025; 93.8% actioned without human review; 97.6% automated decisions confirmed correct. Accessed: 2026-03-17.

[31] Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*, 2025.

[32] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024.

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

[34] Zixuan Wang, Jinghao Shi, Hanzhong Liang, Xiang Shen, Vera Wen, Zhiqian Chen, Yifan Wu, Zhixin Zhang, and Hongyu Xiong. Filter-and-refine: A mllm based cascade system for industrial-scale video content moderation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 873–880, 2025.

[35] xAI. RealWorldQA. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Hugging Face dataset repository. Accessed: 2026-05-03.

[36] 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*, 2025.

[37] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024.