USMI: Benchmarking and Mitigating Unsupported Speech and Music Intrusion in Text Video to Audio Generation

Anonymous ACL submission

Abstract

Text Video to Audio (TV2A) generation can synthesize realistic soundtracks for silent videos, but current models often insert unsupported speech or music that is neither grounded in the visual content nor reliably controllable through text prompts. Standard metrics mainly assess audio quality, semantic alignment, and temporal synchronization, leaving this failure mode largely unmeasured. We define it as Unsupported Speech and Music Insertion (USMI) and study it as a consequence of audiovisual cooccurrence biases, such as off screen speech and background music in training data. To evaluate USMI, we construct USMI-Bench and introduce two direct metrics: USMI@vid, which measures occurrence rate, and USMI@dur, which measures duration ratio. Experiments on mainstream TV2A models show substantial USMI even when conventional metrics remain stable. To mitigate this problem, we propose GUIDE, a training free latent space guidance method that suppresses speech/music activations during sampling while preserving visually grounded audio. Across models and datasets, GUIDE consistently reduces USMI without degrading standard quality or synchronization metrics, and human evaluations confirm improved speech/music faithfulness and overall TV2A quality. This work provides a systematic definition, evaluation, and mitigation framework for USMI in TV2A generation.

1 Introduction

Sound design is central to realism and immersion in films, games, animation, and other multimedia content. Because manual Foley production is expensive and difficult to scale, Text Video to Audio (TV2A) systems aim to synthesize soundtracks directly from visual input. Recent models, including MMAudio (Cheng et al., 2025) and ThinkSound (Liu et al., 2025b), have improved semantic correspondence and temporal synchronization, and are commonly evaluated with metrics such as FD-VGG (Heusel et al., 2017; Hershey et al., 2017), ISC (Salimans et al., 2016), ImageBind score (Girdhar et al., 2023), and DeSync (Iashin et al., 2024).

Despite this progress, current TV2A models exhibit a recurring failure mode that is not directly captured by these metrics: they generate salient speech or music that is unsupported by the visual content. In visually unrelated scenes, such as mechanical actions, household objects, or natural events, a model may insert human voices, narration, singing, or background music that has little connection to what is shown. Because speech and music are perceptually dominant, even a brief insertion can substantially change how listeners interpret the generated soundtrack. A clip may therefore obtain strong distributional quality or temporal alignment scores while remaining semantically unfaithful in a way that human listeners immediately notice.

We define this failure mode as unsupported speech and music insertion (USMI): the generation of additional speech or music on video clips that provide little support for such content. This definition deliberately targets a narrow and measurable phenomenon. We do not claim that all off screen sound is invalid; many real videos naturally contain sounds whose sources are outside the frame. Instead, we ask whether models insert speech or music in a controlled setting where such content has been filtered from the reference signal and is visually unlikely.

This scope is motivated by audiovisual cooccurrence bias in real world video data. Off screen sound is common in such data, and speech and music account for a large portion of off screen content in VGGSound statistics (Figure 1; (Zverev et al., 2025)). These observations suggest that speech and music are not arbitrary target categories, but visible symptoms of priors learned from biased audiovisual cooccurrence. The evaluation challenge is therefore to separate legitimate off screen ambi-

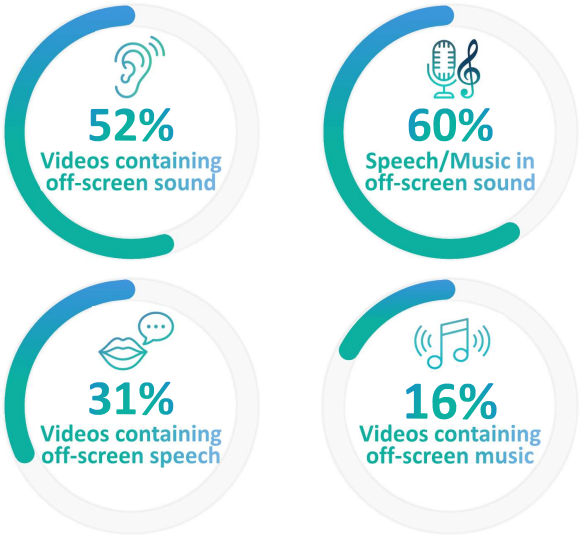

Figure 1: Off screen sound statistics in VGGSound. Speech and music make up a large share of off screen content and motivate our focused study of speech and music insertion.

guity from a diagnostic condition in which speech and music should be unlikely enough that their generation becomes meaningful evidence of model behavior.

To make USMI measurable, we construct USMI-Bench from Kling-Audio-Eval, VGGSound, and AVE. The benchmark combines dataset specific acoustic or annotation based filtering with sampled visual auditing, producing a stricter negative set than a label only protocol. USMI-Bench is designed for precision rather than coverage: the retained clips are not intended to represent all possible videos, but to support a conservative test of whether a model inserts speech or music when such content has been filtered out of the reference signal and visually audited at the dataset level.

We evaluate this behavior with two direct metrics. USMI@vid measures the fraction of generated clips containing detected speech or music, while USMI@dur measures the fraction of generated duration occupied by speech or music. These metrics complement existing quality, semantic, and synchronization scores, which may remain stable even when a generated soundtrack contains salient unsupported content.

Finally, we introduce GUIDE, a training free latent space guidance method for mitigating USMI. GUIDE uses detector feedback on intermediate clean audio estimates to suppress speech and music activations during sampling while encouraging visually grounded alternatives. Our results show that the diagnosed failure mode can be reduced at inference time without retraining the generator.

Overall, this paper reframes visually unsupported audio as a focused diagnosis of unsupported speech and music insertion. Our contributions are threefold: (1) a high precision USMI-Bench that separates off screen ambiguity from measurable speech/music insertion; (2) clip level and duration level USMI metrics validated against human temporal annotations; and (3) GUIDE, a training free inference time guidance method that reduces USMI while preserving standard quality and synchronization metrics. Figure 2 illustrates the evaluation setting and the prior driven insertion behavior studied in this work.

2 Related Work

2.1 Text Video to Audio Generation

Early TV2A work relied on GANs (Chatterjee and Cherian, 2020; Pascual et al., 2017), while recent systems have moved toward diffusion or flow based generation with stronger audiovisual conditioning. Diff-Foley (Luo et al., 2023) uses contrastive audiovisual pretraining to guide latent diffusion. FoleyCrafter (Zhang et al., 2024) introduces semantic and temporal controllers for prompt based generation. MMAudio (Cheng et al., 2025) combines video audio and text audio data to improve semantic coverage, MultiFoley (Chen et al., 2025) supports flexible conditioning with text and audio, and ThinkSound (Liu et al., 2025b) uses multimodal reasoning to plan generation. These models have improved audio quality and synchronization, but their evaluation usually does not isolate whether speech or music is inserted when neither the video nor the reference audio supports it.

2.2 Off screen Audio and Dataset Bias

Real world videos often include off screen sounds. Prior work has therefore studied how to model or evaluate audio that is not visually grounded. VinTAGe (Kushwaha and Tian, 2025) uses text to assist off screen sound generation, Action2Sound (Chen et al., 2024) separates ambient off screen audio from action sounds, and VGGSounder (Zverev et al., 2025) introduces an off screen evaluation dimension. Our work does not claim that off screen audio is always wrong; it isolates speech and music insertion under a filtered negative set setting where the usual off screen explanation is less plausible.

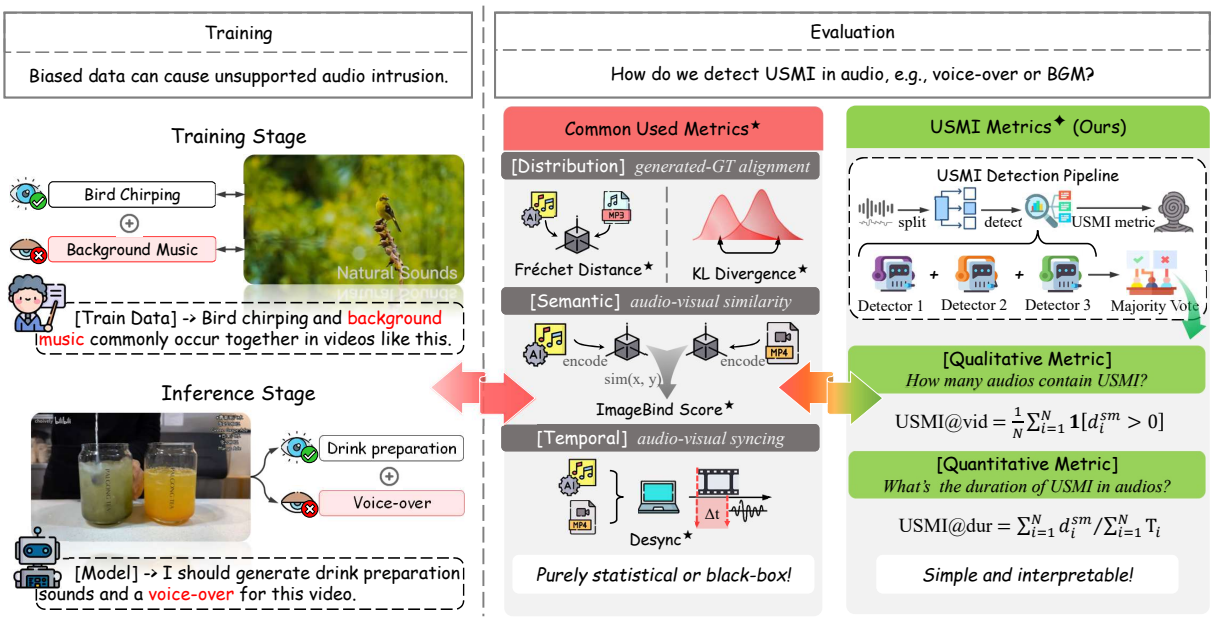

Figure 2: Illustration of unsupported speech and music insertion. Dataset cooccurrence biases can lead a TV2A model to generate voice over or background music on clips selected to lack reference audio or annotation evidence of speech and music.

2.3 Inference Time Guidance

Classifier guidance (Dhariwal and Nichol, 2021) and classifier free guidance (Ho and Salimans, 2022) are widely used to steer diffusion models toward desired attributes. Energy based guidance (Bansal et al., 2024) extends this idea by defining scalar objectives on intermediate predictions and differentiating them with respect to the latent state. We use this mechanism only as a test time mitigation tool, separate from USMI-Bench.

3 USMI-Bench and USMI Measurement

3.1 Problem Scope

We define unsupported speech and music insertion (USMI) as the generation of speech or music on clips selected to lack reference audio or annotation evidence of speech and music, with low estimated visual contamination after manual auditing. This definition is intentionally operational. Because real videos can contain valid off screen audio, USMI is not judged on arbitrary videos; it is judged only on a negative set where dataset labels, reference audio, and sampled visual checks make speech/music support unlikely.

Given a video v, ground truth audio a, and generated audio $\hat{a} = G(v)$, USMI is evaluated only after the sample passes USMI-Bench filtering. Let $D_{\text{sm}}(\cdot)$ denote detected speech and music segments in an audio clip. A generated sample is counted as containing USMI if $D_{\text{sm}}(\hat{a}) \neq \emptyset$. The resulting label should therefore be read as "detected speech/music on a filtered negative set clip," rather than as a universal semantic judgment over all possible off screen events.

3.2 USMI-Bench Construction

USMI-Bench is constructed as a high precision negative set through a dataset specific filtering pipeline. For Kling-Audio-Eval and AVE, we first remove speech and music labeled samples, then apply YAMNet (Ellis et al., 2019), PretrainedSED (Schmid et al., 2025), and PANNs (Kong et al., 2020) to the reference audio with the default threshold 0.30; a clip is removed if any detector reports a contiguous speech or music segment longer than 0.2 s. For VGGSound, we use VGGSounder annotations to remove clips with speech or music evidence, including off screen and background cases. These choices reflect the available supervision in each source dataset: reference audio filtering is used where the original audio is the strongest evidence, while VGGSounder is used where off screen/background annotations provide a more direct signal.

After automatic filtering, we conduct sampled

Dataset	Raw	After label filter	After audio detector filter	Final retained
Kling-Audio-Eval	20,935	15,246	12,836	12,836
VGGSound	15,339	N/A	N/A	4,746
AVE	4,143	2,406	1,080	1,080

Table 1: Reproducible filtering pipeline for USMI-Bench. N/A indicates that the corresponding filtering stage was not applied. VGGSound is filtered with VGGSounder annotations rather than the PANNs and PretrainedSED reference audio detector stage used for Kling-Audio-Eval and AVE.

manual spot checking on the retained clips to estimate remaining visual speech/music contamination. Manual auditing is used as a precision check rather than as an exhaustive relabeling pass, so the benchmark still acknowledges residual ambiguity; this is why USMI is reported as a detector based diagnostic metric and complemented with human preference evaluation. Appendix A gives label groups and detector details, and Appendix B gives the visual audit protocol. Table 1 summarizes the resulting pipeline and retained counts.

3.3 USMI Measurement

For each generated audio clip $\hat{a}_i$, we detect speech and music segments using MV@3, a majority vote ensemble of inaSpeechSegmenter (Doukhan et al., 2018), YAMNet (Ellis et al., 2019), and PretrainedSED (Schmid et al., 2025). We project detector outputs onto a shared 40 ms grid. For detector k, the projected binary mask at frame j is $m_{k,j}(\hat{a}_i) \in \{0, 1\}$, where 1 indicates detected speech or music. MV@3 applies majority voting frame by frame:

$$m_j^{\mathrm{sm}}(\hat{a}_i) = \mathbf{1}\left[\sum_{k=1}^{K} m_{k,j}(\hat{a}_i) \geq \lceil K/2 \rceil\right],$$
$$K = 3, \quad \Delta = 40\mathrm{ms}.$$

Contiguous positive frames in m^{sm} are merged into detected speech and music segments $D_{\mathrm{sm}}(\hat{a}_i)$. This generated audio evaluator is separate from USMI-Bench construction: the benchmark filter operates before generation on source labels, reference audio, or VGGSounder annotations, whereas MV@3 operates after generation on model outputs. Its thresholds are selected on the development split and validated in Sec. 3.4.

USMI is measured at both the clip and duration levels. Let N denote the number of evaluated samples, T_i the total duration of generated sample i, and d_i^{sm} the total duration of detected speech and music in that sample. When detected speech and music segments overlap, we merge overlapping segments before computing d_i^{sm} to avoid double counting. We define

$$\mathrm{USMI@vid} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[d_i^{\mathrm{sm}} > 0],$$
$$\mathrm{USMI@dur} = \frac{\sum_{i=1}^{N} d_i^{\mathrm{sm}}}{\sum_{i=1}^{N} T_i}.$$

USMI@vid measures clip level prevalence, and USMI@dur measures duration level severity. Reporting both matters because a model can affect many clips with short insertions, or fewer clips with long speech/music spans; the two metrics distinguish these failure profiles.

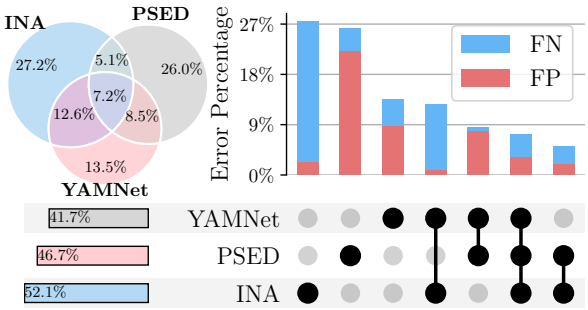

Figure 3: Held out test error overlap across detectors. Errors are largely detector specific, confirming strong decision independence.

3.4 Validation of USMI Measurement

We validate MV@3 on 1,000 generated audio clips sampled from Kling-Audio-Eval, VGGSound, and AVE and produced by MMAudio, ThinkSound, and PrismAudio. Human annotators mark temporal speech and music segments. We split the annotations into 800 development clips and 200 held out test clips, select detector thresholds on the development split, and report precision, recall, F1, and IoU on the held out test split.

Table 2 compares inaSpeechSegmenter, YAMNet, PretrainedSED (Schmid et al., 2025), PANNs (Kong et al., 2020), and MV@3. The individual detectors have complementary precision recall profiles, while MV@3 retains segments only when at least two detectors agree, yielding a precision oriented and robust evaluator. Figure 3 shows limited error overlap across the MV@3 detectors, further supporting majority voting for USMI measurement.

Detector	Label	Thr	Precision	Recall	F1	IoU
ina	speech	-	0.730	0.742	0.736	0.582
yamnet	speech	0.55	0.742	0.747	0.745	0.593
pretrainedsed	speech	0.15	0.774	0.715	0.743	0.591
panns	speech	0.25	0.790	**0.821**	0.805	0.674
MV@3 (ina+yamnet+pretrainedsed)	speech	y=0.55, p=0.15	**0.875**	0.749	**0.807**	**0.677**
ina	music	-	**0.915**	0.585	0.714	0.555
yamnet	music	0.05	0.818	0.816	0.817	0.690
pretrainedsed	music	0.20	0.733	**0.913**	0.813	0.685
panns	music	0.05	0.842	0.896	**0.868**	**0.767**
MV@3 (ina+yamnet+pretrainedsed)	music	y=0.05, p=0.20	0.874	0.822	0.847	0.735

Table 2: Held out test results for speech and music detection after threshold selection on the 800-clip development split. The best result for each metric within each label is shown in bold, and the second best result is underlined.

4 Test Time Steering for USMI Mitigation

USMI-Bench and the metrics above are the main diagnostic contribution. We also include a training free mitigation method to test whether the identified failure mode can be reduced during inference. The method, which we refer to as GUIDE, modifies the sampling trajectory without updating model weights. Its objective is deliberately narrow: suppress speech/music evidence when the current prediction begins to activate those classes, while preserving the generator's ability to produce plausible visible event sounds.

Under the flow matching sampler (Lipman et al., 2022), at a guided sampling step we estimate the clean latent from the current latent as

$$\hat{x}_0 = x_t - t \cdot v_\theta(x_t, t, c).$$

We decode $\hat{x}_0$ into a waveform estimate and feed it to PANNs (Kong et al., 2020), which serves as the differentiable auxiliary classifier for guidance. PANNs is used only to provide a local steering energy during sampling; the final USMI metrics are computed by the separately validated MV@3 evaluator in Sec. 3.3. For each frame ℓ, the speech and music score is the maximum probability over PANNs speech and music subclasses:

$$p_\ell^{sm} = \max_{c \in C_{\text{sp}} \cup C_{\text{mu}}} P_{\text{PANNs}}(c \mid \text{Dec}(\hat{x}_0)_\ell).$$

Let $\mathcal{L}$ denote the set of guidance frames. Frames whose speech and music score exceeds a threshold contribute to a suppression energy,

$$E_{\text{sup}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{ReLU}\big(p_\ell^{sm} - \theta_{\text{sup}}\big).$$

The guidance threshold θ_{sup} is set to the default PANNs threshold 0.30. We keep this value fixed rather than tuning it on USMI test results, so GUIDE remains an inference time control rather than an evaluation specific calibration.

To preserve plausible non speech and non music audio content, GUIDE also selects visually grounded non speech and non music labels from the video and encourages them. Let K denote this selected label set, and let $f_{\ell,k}$ be the PANNs probability of label k at frame ℓ. The promotion energy is

$$E_{\text{pro}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{ReLU}\left(\theta_{\text{pro}} - \max_{k \in K} f_{\ell,k}\right).$$

The total guidance energy is

$$E = E_{\text{sup}} + \lambda E_{\text{pro}}.$$

The corresponding guidance gradient with respect to the current latent is

$$g_t = \nabla_{x_t} E = \left(I - t\frac{\partial v_\theta}{\partial x_t}\right)^\top \nabla_{\hat{x}_0} E.$$

The latent is then updated with the usual reverse Euler denoising step plus normalized energy descent:

$$x_{t-\Delta t} = x_t - \Delta t\, v_\theta(x_t, t, c) - \gamma \frac{g_t}{\|g_t\|_2 + \varepsilon}.$$

The correction penalizes speech and music activations in the intermediate prediction while encouraging visually consistent alternatives. This design makes GUIDE different from prompt negation or CFG adjustment: it intervenes on the emerging audio estimate itself instead of relying only on text conditioning strength. Additional guidance window, robustness, and efficiency analyses are reported in Appendix D.

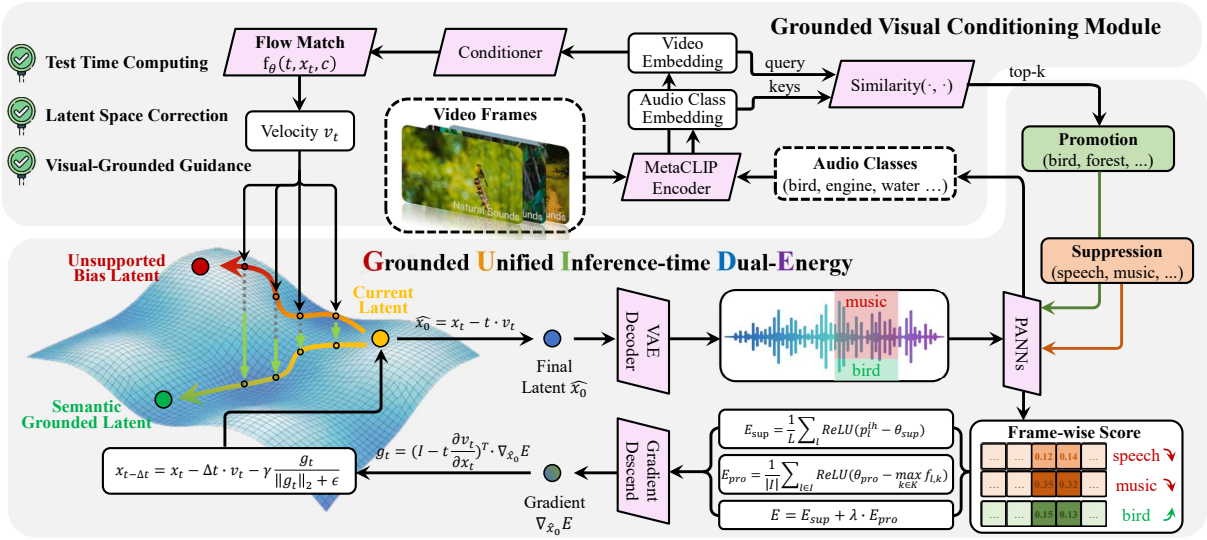

Figure 4: Overview of the test time latent steering procedure used to mitigate USMI.

5 Experiments

This section evaluates the two claims implied by the proposed framework. First, USMI-Bench should reveal speech/music insertion that standard TV2A metrics do not isolate. Second, if the failure is measurable at intermediate sampling states, a targeted test time intervention should reduce it without collapsing ordinary quality metrics. We first report USMI rates across TV2A systems (Section 5.2) and then compare negative prompting, detailed captioning, CFG scale adjustment, and GUIDE across evaluated base models (Section 5.3).

5.1 Experimental Setup

Base models. We evaluate three representative TV2A systems: MMAudio (Cheng et al., 2025), a multimodal framework with flow matching generation and synchronization training; ThinkSound (Liu et al., 2025b), a reasoning based model using Chain of Thought for visual and temporal modeling; and PrismAudio (Liu et al., 2025a), a decomposed chain of thought model with multi dimensional reward optimization. We apply GUIDE to each model in the main USMI-Bench comparison. GUIDE requires no modification to model weights and folds guidance into the existing Euler sampling loop; the more detailed control, human, robustness, and efficiency studies use ThinkSound or PrismAudio as specified in the corresponding tables and appendices.

USMI-Bench. We evaluate on the filtered USMI-Bench splits derived from three datasets: Kling-Audio-Eval (Wang et al., 2025) (12,836 retained clips), VGGSound (Chen et al., 2020) (4,746 retained clips), and AVE (Tian et al., 2018) (1,080 retained clips). Kling-Audio-Eval and AVE are filtered with source labels and reference audio detectors, while VGGSound is filtered with VGGSounder annotations. The original source datasets contain 20,935, 15,339, and 4,143 clips, respectively, but all reported USMI results use the filtered USMI-Bench clips described in Table 1.

Metrics. Unwanted insertion: USMI@vid, USMI@dur. Audio quality: FD-VGG, KL_{PANNs} (Kong et al., 2020) (KL). Semantic/perceptual: ISC, IB-score (Girdhar et al., 2023) (IB). Temporal alignment: DeSync.

5.2 USMI Assessment

We apply the validated USMI metrics to state of the art TV2A models and evaluate whether test time steering reduces unsupported speech and music insertion.

Baseline models exhibit systematic insertion. Table 3 reports results on the filtered USMI-Bench splits. All three models generate speech or music in a substantial fraction of clips where such content is not expected under the evaluation protocol. On Kling-Audio-Eval, USMI@vid ranges from 8.6% for MMAudio to 16.0% for ThinkSound, with USMI@dur reaching 10.6% for ThinkSound. On VGGSound, PrismAudio shows the highest insertion prevalence among the evaluated models. On AVE, all models still show non trivial USMI despite the dataset's frame level audiovisual annotations.

Test time steering reduces USMI. Across the evaluated models, the steering method consis-

Table 3: Evaluation results across USMI-Bench splits. USMI@vid reports clip level prevalence, and USMI@dur reports duration level severity.

Method	Audio Quality Metrics					USMI Metrics	
	FD-VGG ↓	KL-PANNs ↓	ISC ↑	IB-Score ↑	DeSync ↓	USMI@vid ↓	USMI@dur ↓
Kling-Audio-Eval							
MMAudio (Cheng et al., 2025)	10.40	2.50	8.34	0.33	0.63	8.6	3.3
+ GUIDE	10.25	2.47	8.42	0.35	0.64	**6.5** (-24.4%)	**2.2** (-33.3%)
ThinkSound (Liu et al., 2025b)	12.54	2.77	5.52	0.20	0.76	16.0	10.6
+ GUIDE	12.22	2.66	5.57	0.22	0.72	**11.1** (-30.6%)	**7.1** (-33.0%)
PrismAudio (Liu et al., 2025a)	11.36	2.62	6.10	0.29	0.70	13.5	5.4
+ GUIDE	11.34	2.60	6.27	0.29	0.70	**10.5** (-22.2%)	**4.3** (-20.4%)
VGGSound							
MMAudio (Cheng et al., 2025)	6.85	1.81	6.98	0.35	0.58	16.3	6.1
+ GUIDE	6.78	1.79	7.03	0.35	0.58	**10.2** (-37.4%)	**4.5** (-26.2%)
ThinkSound (Liu et al., 2025b)	6.66	2.02	5.74	0.22	0.74	13.6	6.1
+ GUIDE	6.51	1.95	5.79	0.25	0.70	**5.7** (-58.1%)	**2.1** (-65.6%)
PrismAudio (Liu et al., 2025a)	6.17	1.95	6.42	0.31	0.64	19.5	6.7
+ GUIDE	6.06	1.92	6.49	0.31	0.64	**12.3** (-36.9%)	**4.5** (-32.8%)
AVE							
MMAudio (Cheng et al., 2025)	3.23	1.48	6.51	0.39	0.57	13.0	3.1
+ GUIDE	3.24	1.45	6.55	0.39	0.57	**7.8** (-40.0%)	**2.1** (-32.3%)
ThinkSound (Liu et al., 2025b)	8.30	1.94	5.43	0.26	0.70	19.1	7.4
+ GUIDE	8.02	1.90	5.46	0.26	0.70	**11.2** (-41.4%)	**3.5** (-52.7%)
PrismAudio (Liu et al., 2025a)	4.84	1.72	5.89	0.34	0.59	21.4	8.9
+ GUIDE	4.82	1.69	5.94	0.34	0.59	**9.8** (-54.2%)	**3.8** (-57.3%)

tently lowers both clip level and duration level USMI. The reductions are especially large in several high USMI settings, while every USMI-Bench split/model pair shows lower USMI after GUIDE. The method is training free and does not require a second full generation pass, so it can be applied as an inference time correction to existing TV2A systems.

Conventional metrics remain stable. FD, KL, ISC, ImageBind score (Girdhar et al., 2023), and DeSync change only slightly after steering. This supports the main diagnostic point: standard metrics alone do not expose speech and music insertion, and reducing USMI does not necessarily require sacrificing distributional quality, semantic alignment, or temporal synchronization.

5.3 Comparison with Simple Test Time Controls

The diagnostic experiments above show that current TV2A models can insert speech or music even on USMI-Bench. We next evaluate whether this failure mode can be reduced by simple inference time controls under a more challenging setting. Specifically, we construct a 641-video high USMI control subset from Kling-Audio-Eval by selecting categories that show high speech and music insertion across all three evaluated models: Glass, Blender, Kettle whistle, Water tap, and Power tool. We use Kling-Audio-Eval for this comparison because it provides detailed video captions, which are required for the detailed caption intervention.

We compare four classes of test time controls for each base model. *Negative prompt* adds an explicit instruction such as "no speech, no voice, no music, no singing, no background music". *Detailed caption* replaces the default video description with a more explicit non speech and non music caption that enumerates the visible sound sources. *Low CFG* and *high CFG* adjust the classifier free guidance scale to test whether weaker or stronger conditioning suppresses prior driven speech and music. *GUIDE* applies the detector guided latent correction from Sec. 4, which directly penalizes speech and music activations during sampling.

Table 4 reports results on the high USMI Kling-Audio-Eval subset. Across models, negative prompting does not reliably reduce USMI and can even increase insertion, suggesting that prompt level negation is not sufficient to override learned speech and music priors. Detailed captions and lower CFG reduce USMI to some extent, but their effects are model dependent. GUIDE achieves the largest reduction in both USMI@vid and USMI@dur while keeping conventional audio quality, semantic alignment, and synchronization metrics stable.

Method	Audio Quality Metrics					USMI Metrics	
	FD↓	KL↓	ISC↑	IB↑	DeSync↓	USMI@vid↓	USMI@dur↓
MMAudio							
baseline	3.52	1.85	5.01	0.29	0.62	19.1	5.5
+ negative prompt	3.60	1.93	4.78	0.28	0.63	29.4	8.4
+ detailed caption	3.39	1.76	5.03	0.30	0.60	12.1	3.7
+ low CFG	3.49	1.82	4.94	0.29	0.61	16.7	4.8
+ high CFG	3.61	1.91	4.82	0.28	0.63	22.7	6.5
+ GUIDE	3.47	1.80	4.98	0.29	0.61	**8.9**	**2.8**
ThinkSound							
baseline	4.25	2.05	3.32	0.18	0.74	35.6	17.8
+ negative prompt	4.37	2.16	3.09	0.17	0.76	56.6	28.3
+ detailed caption	4.08	1.95	3.30	0.18	0.73	18.6	10.3
+ low CFG	4.20	2.01	3.27	0.18	0.74	30.8	15.2
+ high CFG	4.39	2.12	3.13	0.17	0.76	42.1	21.5
+ GUIDE	4.16	2.00	3.28	0.18	0.73	**13.8**	**7.9**
PrismAudio							
baseline	3.85	1.94	3.67	0.25	0.69	30.0	9.1
+ negative prompt	3.97	2.05	3.44	0.25	0.70	45.6	13.8
+ detailed caption	3.72	1.85	3.65	0.26	0.67	17.0	5.4
+ low CFG	3.80	1.90	3.61	0.26	0.68	25.5	7.7
+ high CFG	3.99	2.01	3.49	0.25	0.70	35.2	10.8
+ GUIDE	3.78	1.89	3.63	0.26	0.68	**12.6**	**4.1**

Table 4: Comparison with simple test time controls on the high USMI Kling-Audio-Eval subset. The 641-video subset contains categories with high speech and music insertion across all three evaluated models: Glass, Blender, Kettle whistle, Water tap, and Power tool. Kling-Audio-Eval is used because it provides detailed video captions required by the detailed caption intervention. Negative prompting and detailed captioning operate at the prompt level, CFG variants adjust the guidance scale, and GUIDE directly suppresses detected speech and music activations during sampling.

5.4 Human Perceptual Evaluation

To verify that the reduction in automatic USMI metrics reflects perceptually meaningful improvements rather than detector specific artifacts, we conduct a human perceptual evaluation on 1,000 clips randomly sampled from the generated outputs of MMAudio, ThinkSound, and PrismAudio. For each sampled clip, we compare the baseline output with its GUIDE counterpart.

Each clip is rated by three independent annotators. Each trial presents the input video and two generated audios in randomized order, and annotators are blind to the system identity. Annotators answered two questions for each pair. First, for **Speech and Music Faithfulness (SM-Faith.)**, they were asked to choose which audio contained fewer unsupported speech or music segments that were inconsistent with the visible video content. Second, for **Overall TV2A Quality (TV2A-Qual.)**, they were asked to choose which generated audio had better overall Text Video to Audio quality, considering audio naturalness, semantic relevance to the video, and temporal synchronization. For each question, annotators could select the baseline output, the GUIDE output, or no difference.

We aggregate annotations at the clip level. A clip is assigned to Baseline or GUIDE if at least two annotators prefer that system. Clips with at least two no difference votes are assigned to the no difference category. Three way disagreement cases, where one annotator selects Baseline, one selects GUIDE, and one selects no difference, are also assigned to the no difference category. We compute 95% confidence intervals by bootstrapping clips. Inter annotator agreement is measured using Fleiss' κ over the three way labels, yielding $\kappa = 0.728$, indicating substantial agreement. Additional protocol details are provided in Appendix C.

Model	Criterion	Baseline	GUIDE	No Difference
MMAudio	SM-Faith.	12.1 [8.6, 15.6]	**58.7 [53.4, 64.0]**	29.2 [24.3, 34.1]
	TV2A-Qual.	16.2 [12.2, 20.2]	**67.6 [62.6, 72.6]**	16.2 [12.2, 20.2]
ThinkSound	SM-Faith.	8.9 [5.8, 12.0]	**66.8 [61.7, 71.9]**	24.3 [19.7, 28.9]
	TV2A-Qual.	13.6 [9.9, 17.3]	**73.1 [68.3, 77.9]**	13.3 [9.7, 16.9]
PrismAudio	SM-Faith.	10.8 [7.5, 14.1]	**64.1 [58.9, 69.3]**	25.1 [20.4, 29.8]
	TV2A-Qual.	14.6 [10.8, 18.4]	**72.9 [68.1, 77.7]**	12.5 [8.9, 16.1]

Table 5: Human perceptual evaluation by model. Values are percentages, and brackets denote estimated 95% confidence intervals. SM-Faith. measures which audio contains fewer unsupported speech and music segments inconsistent with the visible video. TV2A-Qual. measures which audio has better overall Text Video to Audio quality.

Table 5 shows that GUIDE receives more preferences than the baseline in the reported pairwise evaluation, with non trivial no difference rates. Under SM-Faith., annotators prefer GUIDE over the corresponding baseline, indicating that GUIDE's USMI reduction is perceptually meaningful rather than a detector specific artifact. Under TV2A-Qual., GUIDE also receives more preferences across models, suggesting that reducing unsupported speech and music does not come at the cost of degraded perceived TV2A quality in this evaluation.

6 Conclusion

We studied unsupported speech and music insertion in Text Video to Audio generation through USMI-Bench and two direct USMI metrics. Experiments show that current TV2A models can still insert speech or music at substantial rates while conventional quality and synchronization metrics remain stable. A latent space steering method reduces this failure mode during sampling. Future work should extend this diagnostic framework beyond speech and music while preserving the same level of USMI-Bench precision.

7 Limitations

This work deliberately narrows the evaluation target to unsupported speech and music insertion, so it does not cover other forms of audio hallucination, such as unsupported environmental sounds, incorrect sound timing, or mismatched acoustic material. USMI-Bench may still contain residual false negatives or ambiguous visual contexts despite conservative filtering and auditing. Our experiments cover representative open TV2A models and datasets, but the measured USMI rates and steering effectiveness may vary for proprietary systems, different training distributions, longer videos, or prompts with richer semantic context. GUIDE reduces USMI at inference time, but it does not directly improve the underlying model's learned audiovisual priors and may require additional validation with different samplers, guidance schedules, or downstream editing pipelines.

8 Information About Use of AI Assistants

AI assistants were used to support language polishing, organization of manuscript text, and LaTeX editing. The research ideas, USMI-Bench design, experiments, data analysis, claims, and final manuscript decisions were made and verified by the authors. The authors take full responsibility for the content of this paper.

References

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Universal guidance for diffusion models. In *International Conference on Learning Representations*, volume 2024, pages 51304–51323.

Moitreya Chatterjee and Anoop Cherian. 2020. Sound2sight: Generating visual dynamics from sound and context. In *European Conference on Computer Vision*, pages 701–719. Springer.

Changan Chen, Puyuan Peng, Ami Baid, Zihui Xue, Wei-Ning Hsu, David Harwath, and Kristen Grauman. 2024. Action2sound: Ambient-aware generation of action sounds from egocentric videos. In *European Conference on Computer Vision*, pages 277–295. Springer.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. 2025. Video-guided foley sound generation with multimodal controls. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18770–18781.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. 2025. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28901–28911.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

David Doukhan, Eliott Lechapt, Marc Evrard, and Jean Carrive. 2018. Ina's mirex 2018 music and speech detection system. In *Music Information Retrieval Evaluation eXchange (MIREX 2018)*.

Dan Ellis, Shawn Hershey, Ron J. Weiss, Kevin Wilson, and J. R. Moore. 2019. Yamnet: Yet another mobilenet for audio event classification. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Malcolm Slaney, Ron J Weiss, and 1 others. 2017. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. 2024. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329. IEEE.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Saksham Singh Kushwaha and Yapeng Tian. 2025. Vintage: Joint video and text conditioning for holistic audio generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13529–13539.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Huadai Liu, Kaicheng Luo, Wen Wang, Qian Chen, Peiwen Sun, Rongjie Huang, Xiangang Li, Jieping Ye, and Wei Xue. 2025a. Prismaudio: Decomposed chain-of-thoughts and multi-dimensional rewards for video-to-audio generation. *arXiv preprint arXiv:2511.18833*.

Huadai Liu, Jialei Wang, Kaicheng Luo, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. 2025b. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing. *arXiv preprint arXiv:2506.21448*.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. 2023. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:48855–48876.

Santiago Pascual, Antonio Bonafonte, and Joan Serra. 2017. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Florian Schmid, Tobias Morocutti, Francesco Foscarin, Jan Schlüter, Paul Primus, and Gerhard Widmer. 2025. Effective pre-training of audio transformers for sound event detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *ECCV*.

Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, and 1 others. 2025. Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. 2024. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*.

Daniil Zverev, Thaddäus Wiedemer, Ameya Prabhu, Matthias Bethge, Wieland Brendel, and A Koepke. 2025. Vggsounder: Audio-visual evaluations for foundation models. *arXiv preprint arXiv:2508.08237*.

A USMI-Bench Filtering Details

This appendix specifies the label filters, detector target groups, thresholds, and temporal granularity used in USMI-Bench construction. Tables 9 and 10 provide the complete label groups used for reproducibility.

USMI-Bench construction and generated audio evaluation use separate threshold policies. For Kling-Audio-Eval and AVE USMI-Bench construction, YAMNet (Ellis et al., 2019), PretrainedSED (Schmid et al., 2025), and PANNs (Kong et al., 2020) are applied to ground truth audio with the default threshold 0.30; after projecting detector outputs onto a shared 40 ms grid, a clip is removed if any detector produces a contiguous speech or music segment longer than 0.2 s. For USMI evaluation, the MV@3 thresholds are selected on the 800-clip generated audio development split and reported in Table 2. For GUIDE, the PANNs guidance threshold is fixed at the default threshold 0.30. This does not create circular use of detectors: USMI-Bench filtering is performed on ground truth/reference audio before generation, whereas USMI evaluation is performed on generated audio, which is a different distribution and a later stage of the protocol.

B Manual Visual Spot-Checking Criteria

Manual visual spot checking uses a conservative checklist to identify obvious visual evidence that would reasonably support speech or music generation. The audit is performed after automatic label/annotation and audio detector filtering. We randomly sample 600 retained clips from Kling-Audio-Eval, 200 retained clips from VGGSound, and 50 retained clips from AVE. Reviewers inspect the video frames and mark whether the clip contains visible speaking people, singers, musicians, musical instruments, stages, concerts, performances, microphones, loudspeakers, music videos, dance performances, interviews, lectures, presentations,

Dataset	Manual audit sample size	Suspicious clips	Suspicious rate
Kling-Audio-Eval	600	12	2.0%
VGGSound	200	2	1.0%
AVE	50	2	4.0%
Total	**850**	16	1.9%

Table 6: Manual visual spot checking protocol for the retained USMI-Bench. Reviewers flag obvious speech and music related visual cues; ambiguous cases are counted as visually suspicious.

or other explicit speech and music related cues. Ambiguous samples are counted as visually suspicious. Table 6 summarizes the audit protocol.

C Human Evaluation Protocol Details

For the human perceptual evaluation in Table 5, annotators are shown the input video together with two generated audio tracks: the baseline output and the corresponding GUIDE output. The presentation order of the two audio tracks is randomized independently for each trial, and system names are hidden from annotators. Annotators make two separate three way judgments for each clip: Speech and Music Faithfulness (SM-Faith.) and Overall TV2A Quality (TV2A-Qual.). The available labels are Baseline preferred, GUIDE preferred, and No Difference.

The final clip level label is obtained by majority vote over the three annotators. A two vote majority for Baseline, GUIDE, or No Difference assigns the clip to that category. If all three labels occur once, the clip is treated as No Difference, since there is no stable preference between the two systems. Confidence intervals are obtained by bootstrapping clips, and inter annotator agreement is measured with Fleiss' κ over the three labels.

D Supplementary Experiments for GUIDE

Unless otherwise specified, GUIDE uses the default PANNs speech and music detection threshold 0.30 for θ_{sup} and sets the guidance strength to $\gamma = 1.0$. These settings are kept fixed across the supplementary experiments below.

D.1 Steering Window

We examine when unsupported speech and music become fixed during generation. Starting from partially corrupted ground truth audio latents, we let the model continue denoising from different steps and measure the resulting USMI@dur. In the GUIDE experiments, guidance is applied at

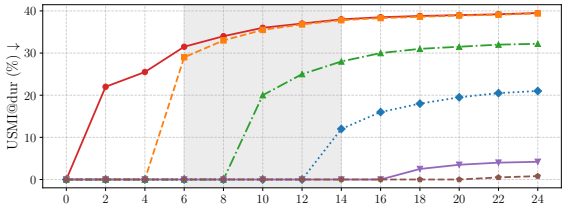

Figure 5: Intermediate sampling steps provide the effective window for reducing USMI before speech and music content becomes fixed in the trajectory.

sampling steps $[8, 10, 12, 14]$ out of 24 total steps. As shown in Figure 5, speech and music insertion rises sharply in the middle of the sampling process and then stabilizes. This suggests that intermediate steps are the most useful intervention window: the output is structured enough for a detector to provide a meaningful signal, but not yet so fixed that correction becomes ineffective.

D.2 Robustness to Detector Errors

We analyze whether detector level errors propagate through the pipeline and compromise USMI metric reliability or GUIDE's effectiveness. Starting from validated MV masks, we simulate detector degradation by introducing borderline false positives (+FP) and false negatives (+FN) at a 20% rate, and apply both simultaneously (+FP +FN). Degraded masks serve as both the guidance signal and the evaluation basis, representing the worst case propagation scenario. This experiment uses the same ThinkSound control subset as Table 4.

As shown in Table 7, all perturbed conditions retain substantial USMI reduction relative to the ThinkSound baseline. The +FP condition causes GUIDE to apply suppression on borderline non hallucinated frames, but because their speech and music probabilities are near the threshold θ_{sup}, the ReLU gated energy is small and audio quality is unaffected. The +FN condition removes low confidence detections that contribute little suppression signal, so USMI reduction is largely preserved. Even under +FP +FN, USMI@vid and USMI@dur are reduced by 58% and 59%, and all quality metrics remain stable. Error propagation is therefore bounded and does not undermine the system's core reliability.

D.3 Inference Efficiency

GUIDE folds guidance into the existing Euler sampling loop, adding VAE decode and PANNs forward/backward passes at 4 guided steps out of 24

Method	FD↓	KL↓	ISC↑	IB↑	DeSync↓	USMI@	
						vid↓	dur↓
ThinkSound	4.249	2.046	**3.318**	0.175	0.744	35.6	17.8
GUIDE	**3.899**	**1.881**	3.203	**0.183**	**0.731**	13.8	7.9
+FP(20%)	3.921	1.903	3.219	0.182	0.734	19.0	9.3
+FN(20%)	3.942	1.917	3.231	0.181	0.737	**13.5**	**7.2**
+FP +FN	3.958	1.928	3.245	0.180	0.739	14.8	7.4

Table 7: Robustness of GUIDE to simulated detector errors on the ThinkSound control subset (+FP/+FN on borderline detections). KL and ISC are PANNs based. All conditions retain substantial USMI reduction.

total sampling steps, with Stage B (promote label selection) executed entirely on CPU. Table 8 reports wall clock time and peak GPU memory for each base model with and without GUIDE, measured on a single RTX 5880 Ada 48G GPU with batch size 1 on Kling-Audio-Eval. GUIDE incurs approximately **2.3 to 2.7**× time overhead due to the gradient computation at each guided step, and increases peak VRAM to 24.3 GB for ThinkSound and 15.0 GB for PrismAudio to accommodate the VAE decoder and PANNs backward pass.

For the full GUIDE update, the latent space guidance direction is

$$g_t = \left(I - t\frac{\partial v_\theta}{\partial x_t}\right)^\top \nabla_{\hat{x}_0}\mathcal{E},$$

where $\mathcal{E}$ denotes the detector energy computed from the decoded clean audio estimate. In late sampling steps, $t \approx 0$, so the Jacobian factor $I - t\,\partial v_\theta/\partial x_t$ is close to I. We therefore use a faster approximation that ignores the velocity Jacobian:

$$g_t \approx \nabla_{\hat{x}_0}\mathcal{E}.$$

Because the update uses normalized gradients, the approximation is applied after normalization,

$$\frac{g_t}{\|g_t\|_2 + \varepsilon} \approx \frac{\nabla_{\hat{x}_0}\mathcal{E}}{\|\nabla_{\hat{x}_0}\mathcal{E}\|_2 + \varepsilon}.$$

Empirically, late step guidance directions have high cosine similarity under this approximation, so back propagation through v_θ can be skipped after step 16 of 24. Applying this approximation reduces the time overhead to approximately **1.7 to 1.8**× and lowers peak VRAM to 15.6 GB for ThinkSound and 12.4 GB for PrismAudio, with minimal impact on USMI reduction.

Base Model	Method	Efficiency Metrics	
		Time↓	VRAM↓
ThinkSound	Baseline	1.20	10.1
	+ GUIDE	2.72 (2.3×)	24.3
	+ GUIDE (approx.)	2.13 (1.8×)	15.6
PrismAudio	Baseline	1.24	5.4
	+ GUIDE	3.30 (2.7×)	15.0
	+ GUIDE (approx.)	2.14 (1.7×)	12.4

Table 8: Inference time and peak VRAM. Time is reported in ***seconds per clip***; VRAM is reported in ***GB***. Batch size is 1 on RTX 5880 Ada 48G. *Approx.* ignores the velocity Jacobian in the chain rule guidance gradient after step 16.

Dataset	Speech labels removed	Music labels removed	Filtering note
Kling-Audio-Eval	*Human sounds/* directory	*Music/* directory	Directory level source label filtering before reference audio detector filtering.
VGGSound	child speech, kid speaking; female speech, woman speaking; male speech, man speaking; police radio chatter; people whispering; parrot talking; people babbling; children shouting; people screaming; baby babbling; people sobbing; people coughing; people hiccup; people humming; people battle cry; people giggling; people nose blowing; people gargling	female singing; male singing; beat boxing; child singing; yodelling; singing choir; rapping; people humming; orchestra; playing accordion; playing acoustic guitar; playing banjo; playing bass guitar; playing bassoon; playing cello; playing cymbal; playing didgeridoo; playing drum kit; playing electric guitar; playing flute; playing harp; playing marimba, xylophone; playing piano; playing saxophone; playing trombone; playing violin, fiddle; singing bowl; playing clarinet; playing hammond organ; playing tabla; playing glockenspiel; playing harpsichord; playing theremin; playing snare drum; playing erhu; playing bagpipes; playing steel guitar, slide guitar; playing bongo; playing synthesizer; playing ukulele; playing cornet; playing vibraphone; playing french horn; playing bass drum; playing sitar; playing electronic organ; playing trumpet; playing harmonica; playing mandolin; playing tambourine; playing double bass; tapping guitar; playing timpani; playing timbales; playing gong; playing oboe; playing zither; playing steelpan; playing guiro; playing congas; playing tympani; playing bugle; playing djembe; wind chime; disc scratching; metronome; playing shofar; playing tuning fork; playing washboard; playing castanets	VGGSounder annotation filtering removes on screen, off screen, background, and metadata speech and music evidence.
AVE	Male speech, man speaking; Female speech, woman speaking; Baby cry, infant cry	Accordion; Acoustic guitar; Banjo; Flute; Mandolin; Shofar; Ukulele; Violin, fiddle	Class label filtering before reference audio detector filtering.

Table 9: Source label filters used for USMI-Bench construction. The complete speech and music label groups are embedded as table entries to make the negative set filtering protocol reproducible.

Detector	Ontology or native output	Speech target labels	Music target labels	Threshold use
inaSpeech Segmenter	Native segment categories	speech	music	No tunable confidence threshold.
YAMNet and PANNs	AudioSet 527-class ontology (Gemmeke et al., 2017)	speech; male speech, man speaking; female speech, woman speaking; child speech, kid speaking; conversation; narration, monologue; babbling; speech synthesizer; shout; bellow; whoop; yell; battle cry; children shouting; screaming; whispering; laughter; baby laughter; giggle; snicker; belly laugh; chuckle, chortle; crying, sobbing; baby cry, infant cry; whimper; wail, moan; sigh; hubbub, speech noise, speech babble; chatter	singing; male singing; female singing; choir; yodeling; chant; mantra; child singing; synthetic singing; rapping; humming; whistling; beatboxing; music; musical instrument; plucked string instrument; guitar; electric guitar; bass guitar; acoustic guitar; steel guitar, slide guitar; tapping (guitar technique); strum; banjo; sitar; mandolin; zither; ukulele; keyboard (musical); piano; electric piano; organ; electronic organ; hammond organ; synthesizer; sampler; harpsichord; percussion; drum kit; drum machine; drum; snare drum; rimshot; drum roll; bass drum; timpani; tabla; cymbal; hi-hat; wood block; tambourine; rattle (instrument); maraca; gong; tubular bells; mallet percussion; marimba, xylophone; glockenspiel; vibraphone; steelpan; orchestra; brass instrument; french horn; trumpet; trombone; bowed string instrument; string section; violin, fiddle; pizzicato; cello; double bass; wind instrument, woodwind instrument; flute; saxophone; clarinet; harp; bell; church bell; jingle bell; bicycle bell; tuning fork; chime; wind chime; change ringing (campanology); harmonica; accordion; bagpipes; didgeridoo; shofar; theremin; singing bowl; scratching (performance technique); pop music; hip hop music; rock music; heavy metal; punk rock; grunge; progressive rock; rock and roll; psychedelic rock; rhythm and blues; soul music; reggae; country; swing music; bluegrass; funk; folk music; middle eastern music; jazz; disco; classical music; opera; electronic music; house music; techno; dubstep; drum and bass; electronica; electronic dance music; ambient music; trance music; music of latin america; salsa music; flamenco; blues; music for children; new-age music; vocal music; a capella; music of africa; afrobeat; christian music; gospel music; music of asia; carnatic music; music of bollywood; ska; traditional music; independent music; song; background music; theme music; jingle (music); soundtrack music; lullaby; video game music; christmas music; dance music; wedding music; happy music; funny music; sad music; tender music; exciting music; angry music; scary music	GT audio filtering uses default threshold 0.30 and removes clips with any contiguous positive segment longer than 0.2 s. For generated audio evaluation, YAMNet uses F1 optimized thresholds selected on the 800-clip development split and reported in Table 2. For GUIDE, PANNs uses the default threshold 0.30.
PretrainedSED	Configured AudioSet target groups	Same speech target group as the AudioSet row above.	Same music target group as the AudioSet row above.	Same threshold policy as above: default 0.30 for GT audio filtering, and F1 optimized thresholds from the 800-clip generated audio development split for USMI evaluation.

Table 10: Detector target label groups used for USMI-Bench audio filtering, USMI evaluation, and GUIDE guidance. YAMNet and PANNs share the same AudioSet target groups; PretrainedSED uses the corresponding configured target groups.