AV-RIFT: TASK-AGNOSTIC TRANSFERABLE ADVERSARIAL ATTACKS ON AUDIO-VISUAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-visual models integrate auditory and visual cues to provide a comprehensive understanding of real-world scenes and have been applied to diverse tasks (e.g., event localization and question answering), largely benefiting from pretrained unimodal encoders that provide well-generalized representations. However, the open accessibility of these encoders also exposes a shared adversarial attack surface across downstream models, which remains unexplored by existing audio-visual attacks. This paper, for the first time, explores this threat by proposing **AV-RIFT**, a transferable audio-visual attack that solely utilizes publicly available encoders to mislead diverse downstream models. It jointly perturbs audio and video by disrupting semantic relationships both within and across modalities. Specifically, since heterogeneous audio-visual tasks commonly revolve around the semantics of sounding events shared by different instances, merely disrupting instance-level features on pretrained encoders does not necessarily alter the perceived event. AV-RIFT therefore introduces an intra-modal semantics-aware contrastive attack that steers each modality's representations away from semantically similar references and toward dissimilar ones. Moreover, to exploit the semantic consistency between audio and visual streams overlooked by independent unimodal perturbations, it further introduces an inter-modal misalignment attack that reduces the similarity between synchronized audio and visual representations in a projection space learned to bridge independently pretrained encoders. By exploiting the event-centric semantics and temporal synchronization inherent in audio-visual data, AV-RIFT is tailored to audio-visual models rather than being a direct adaptation of attacks designed for other modalities. Experiments on 5 audio-visual tasks and 20 models show that AV-RIFT consistently achieves state-of-the-art attack transferability and remains effective under existing defenses.

1 INTRODUCTION

Audio-visual models integrate auditory and visual cues to provide a comprehensive understanding of real-world scenes, and have achieved remarkable success across a wide range of tasks, including Audio-Visual Event Localization (AVEL) (Tian et al., 2018), Audio-Visual Segmentation (AVS) (Zhou et al., 2022b), Audio-Visual Event Recognition (AVER) (Cristani et al., 2007), Audio-Visual Retrieval (AVR) (Zeng et al., 2020), and Audio-Visual Question Answering (AVQA) (Li et al., 2022a). More recently, general audio-visual models (Du et al., 2025; Tang et al., 2026; Vyas et al., 2026) further unify multiple tasks within a single framework. Despite substantial differences in architecture and output, these models largely follow a common paradigm (Wei et al., 2022a): pretrained modality-specific encoders with well-generalized representations extract audio and visual features, which are then fused or compared by task-specific modules to produce predictions.

Like their unimodal counterparts, audio-visual models are also vulnerable to adversarial perturbations (Wen et al., 2025), raising serious concerns about their reliability. Existing studies mainly adapt unimodal attacks (e.g., PGD (Madry et al., 2017)) by perturbing one or both modalities independently (Tian & Xu, 2021; Yang et al., 2021; Li et al., 2022b; Yang et al., 2024). Such attacks do not exploit the relationships between the two modalities. Zhang et al. (2025) further propose TMA, which exploits temporal redundancy and cross-modal correlations to optimize perturbations. As shown in Fig. 1(a)–(b), these audio-visual attacks assume access to the task-specific model and are exclusively designed for AVER, leaving other audio-visual tasks largely unexplored.

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

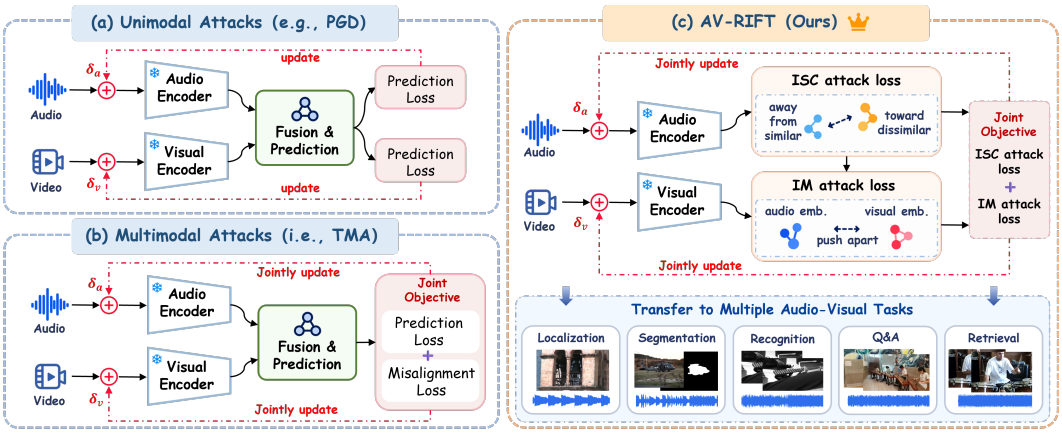


Figure 1: **Comparison of audio-visual attack paradigms.** (a) Unimodal attacks independently perturb the audio and visual modalities. (b) TMA exploits inter-modal relations but remains tied to task-specific prediction objectives. (c) AV-RIFT exploits intra-modal semantic disruption and inter-modal misalignment in surrogate representation spaces, enabling task-agnostic transfer across heterogeneous audio-visual tasks.

In practice, however, this assumption rarely holds, as real-world audio-visual models are typically deployed as black-box services (Ye et al., 2025) that conceal their architectures and gradients from attackers. Nevertheless, since these models are commonly built upon publicly accessible pretrained encoders, a natural question arises: *can audio-visual perturbations crafted only on public encoders mislead diverse downstream models across heterogeneous tasks, posing a more practical threat?*

Evidence from unimodal domains suggests that this is plausible: attacks on pretrained encoders have proven transferable in the video (Wei et al., 2022b; Lu et al., 2026) and audio (Hussain et al., 2026) domains. However, directly combining them for audio-visual inputs faces two limitations. First, heterogeneous audio-visual tasks commonly revolve around the semantics of sounding events, which are shared by different instances (Morgado et al., 2021). Since these attacks merely push each adversarial feature away from its own clean counterpart at the instance level, the perturbed representation may remain within the semantic neighborhood of the original event, leaving the perceived event unchanged. Second, perturbing each modality independently overlooks the semantic consistency between synchronized audio and visual streams, which plays a key role in downstream predictions (Ye et al., 2026).

To bridge this gap, we propose **AV-RIFT**, a task-agnostic transferable audio-visual attack that jointly perturbs audio and video by disrupting semantic relationships both within and across modalities. As shown in Fig. 1(c), AV-RIFT combines two complementary objectives. **1)** To alter the perceived event rather than merely deviating at the instance level, the **intra-modal semantics-aware contrastive attack** defines contrastive relations based on semantic relatedness. Using clean features of other examples as references, it pushes each modality's adversarial representation out of the semantic neighborhood of the original event and toward those of semantically dissimilar events. As event semantics are shared across tasks, this shift can mislead heterogeneous downstream models without relying on any task objective. **2)** Beyond unimodal semantics, audio-visual predictions also depend on the agreement between modalities, where one modality can corroborate or compensate for the other. The **inter-modal misalignment attack** therefore breaks the correspondence between synchronized audio and visual representations, turning complementary cues into conflicting evidence for cross-modal fusion. To this end, it measures this correspondence in a projection space that aligns independently pretrained encoders, which is learned from auxiliary paired data and further couples the optimization of the audio and visual perturbations. By exploiting the event-centric semantics and temporal synchronization inherent in audio-visual data, AV-RIFT is tailored to audio-visual models rather than being a direct adaptation of attacks designed for other modalities.

We evaluate AV-RIFT on 5 audio-visual tasks and 20 task-specific and general models, including audio-visual large language models. Our method consistently achieves the strongest transferability,

improving ASR by 12.87% on average over the strongest baseline on task-specific models and by 13.00% on general audio-visual models. On the AVER task, it also outperforms the task-specific TMA scheme, and it remains effective under different defenses.

Our contributions are summarized as follows:

1. We propose AV-RIFT, a task-agnostic transferable audio-visual attack that requires only public surrogate encoders. To the best of our knowledge, it is the first attack to expose the shared attack surface of public encoders, achieving transfer across heterogeneous tasks by exploiting the semantic relationships inherent in audio-visual data.

2. We design an intra-modal semantics-aware contrastive attack that shifts adversarial representations out of the original event's semantic neighborhood, altering the perceived event in a task-agnostic manner, and an inter-modal misalignment attack that turns complementary audio-visual cues into conflicting evidence, using lightweight projections that align independently pretrained encoders and couple the optimization of both perturbations.

3. Experiments on 5 tasks and 20 models under different attack settings show that AV-RIFT consistently outperforms existing attacks and remains effective under defenses.

2 Related Work

Audio-Visual Learning. Audio-visual learning (Zhu et al., 2021) covers diverse tasks with distinct output structures, including temporal event localization (Tian et al., 2018), pixel-level sounding object segmentation (Zhou et al., 2022b), event recognition (Cristani et al., 2007), cross-modal retrieval (Zeng et al., 2020), and question answering (Li et al., 2022a), with substantial progress from task-specific models (He et al., 2024; Chen et al., 2024; Kim et al., 2025). Recent unified models (Vyas et al., 2026), including audio-visual large language models (Du et al., 2025; Tang et al., 2026), further support multiple tasks within a single framework. Despite their diversity, these models generally encode audio and visual inputs with pretrained modality-specific encoders and integrate them through cross-modal fusion (Wei et al., 2022a). This common reliance may introduce a shared attack surface, through which a single attack design can target multiple tasks.

Adversarial Attacks. Adversarial vulnerability has been extensively studied in unimodal domains (Chakraborty et al., 2018; Wang et al., 2026), ranging from gradient-based attacks such as PGD (Madry et al., 2017) and their transferability-enhanced variants (Dong et al., 2018; Xie et al., 2019) to audio attacks that account for the temporal structure (Abdoli et al., 2019) and perceptual characteristics (Qin et al., 2019) of audio signals. Correspondingly, defenses such as adversarial training (Bai et al., 2021) and input transformations (Xu et al., 2017; Dziugaite et al., 2016; Hussain et al., 2021; Chen et al., 2022) have been developed. In comparison, the adversarial vulnerability of audio-visual models remains underexplored (Wen et al., 2025). Existing studies mainly focus on improving robustness (Tian & Xu, 2021; Yang et al., 2021; Li et al., 2022b; Yang et al., 2024), where adversarial examples are constructed by applying unimodal attacks to one or both modalities independently. Zhang et al. (2025) proposed TMA, which exploits the temporal and cross-modal properties of audio-visual data, but it relies on the classification objective of an AVER model. In contrast, AV-RIFT exploits the shared attack surface across audio-visual tasks and models, without accessing any downstream model.

Transfer Attacks on Pretrained Encoders. Attacking without accessing downstream models is closely related to transfer-based attacks, which craft adversarial examples on surrogate models to mislead unseen victim models (Wang et al., 2026), but typically assume that the surrogate and victim share the same task. To relax this assumption, recent works attack the representation space of pretrained encoders, showing that the resulting perturbations transfer to diverse downstream models without task knowledge in the video (Wei et al., 2022b; Huang et al., 2026; Lu et al., 2026) and audio (Hussain et al., 2026) domains. These findings support the shared attack surface of audio-visual models discussed above. Beyond unimodal domains, multimodal attacks on pretrained encoders have been mainly explored for vision-language models (Lu et al., 2023; Wang et al., 2024), which typically exploit the image–text embedding space aligned by contrastive pretraining. However, none of these attacks is designed for the audio-visual domain.

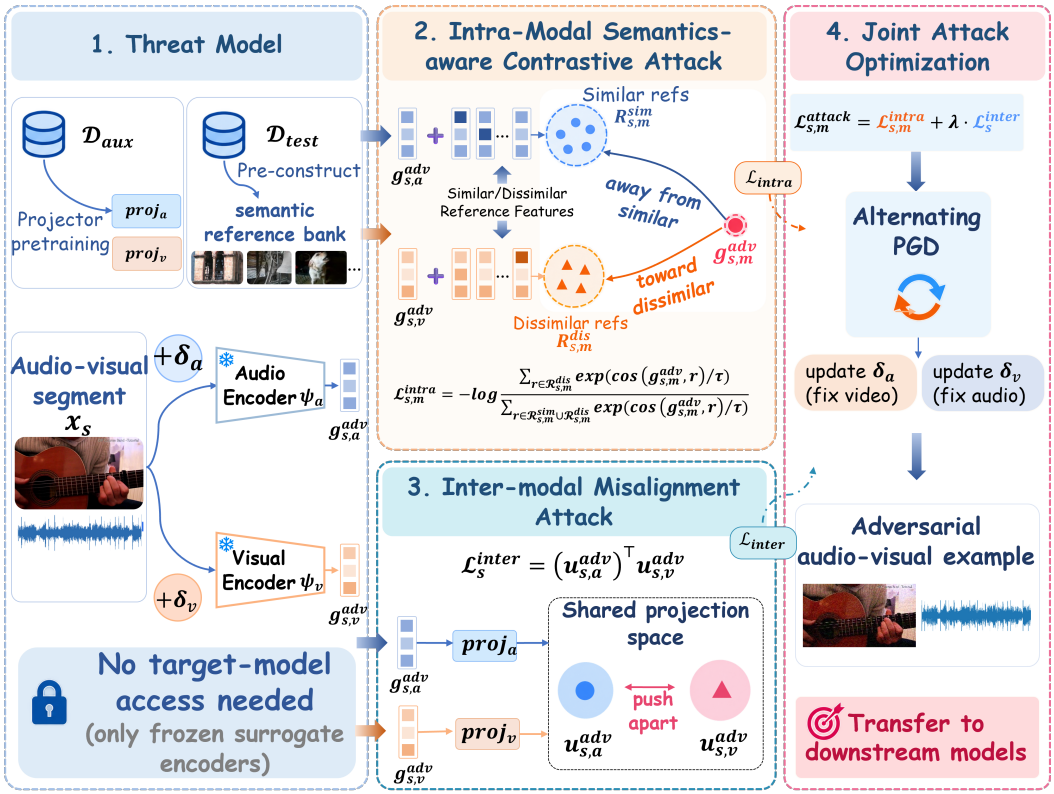


Figure 2: **Overview of AV-RIFT.** Using only frozen surrogate encoders, AV-RIFT combines an intra-modal semantics-aware contrastive attack and an inter-modal misalignment attack, which are jointly optimized with alternating PGD to generate transferable audio-visual adversarial examples.

3 AV-RIFT

In this section, we first introduce the threat model and notations (Section 3.1). We then present the two components of AV-RIFT, namely the intra-modal semantics-aware contrastive attack (Section 3.2) and the inter-modal misalignment attack (Section 3.3), followed by their joint optimization (Section 3.4). An overview is shown in Figure 2.

3.1 Preliminaries

Threat Model. We consider untargeted transfer attacks without access to task-specific models, i.e., the adversary can neither query nor differentiate through the downstream encoders or prediction modules. The adversary has access only to the complete labeled test set $\mathcal{D}_{\text{test}}$, whose labels are used solely to group semantically related references and can be replaced by feature-space neighborhoods (Section 3.2), publicly available surrogate audio and visual encoders ψ_a, ψ_v, and an auxiliary paired audio-visual dataset $\mathcal{D}_{\text{aux}}$ for learning the modality projections. The surrogate encoders are not required to match the encoders of the downstream models.

Notations. We divide each audio-visual input X into S synchronized segments, and let $x_{s,a}$ and $x_{s,v}$ denote the audio waveform and sampled video frame of segment s, respectively. Taking AVEL as a running example, a downstream model predicts

$$\widehat{y}_s = \mathcal{F}_\theta(f_{s,a}, f_{s,v}) \tag{1}$$

where $f_{s,m} = \phi_m(x_{s,m})$, $m \in \{a, v\}$, denotes the feature extracted by the model's audio or visual encoder ϕ_m, and $\mathcal{F}_\theta$ is the task-specific fusion and prediction module; both are unknown to the adversary. Other audio-visual tasks follow the same paradigm, differing mainly in output structure and temporal granularity.

We perform a segment-wise feature attack by optimizing an input perturbation $\delta_{s,m}$ for each segment s and modality $m \in \{a, v\}$. The perturbed inputs are $x^{\text{adv}}_{s,m} = x_{s,m} + \delta_{s,m}$, subject to $\|\delta_{s,m}\|_\infty \leq \epsilon_m$ and valid input ranges. The frozen surrogates extract clean features $g_{s,m} = \psi_m(x_{s,m})$ and adversarial features $g^{\text{adv}}_{s,m} = \psi_m(x^{\text{adv}}_{s,m})$. The attack objectives operate on these surrogate features, while optimization is performed only over the input perturbations.

3.2 Intra-Modal Semantics-aware Contrastive Attack

Existing attacks on pretrained encoders (Wei et al., 2022b; Hussain et al., 2026) typically push each adversarial feature away from its own clean counterpart, e.g., by minimizing $\cos(g^{\text{adv}}_{s,m}, g_{s,m})$. However, since different instances can share the same event semantics, such instance-level deviation does not control the *semantic direction* of the displacement: the adversarial feature may still lie close to other examples of the same event, allowing downstream models to retain correct predictions. We therefore define contrastive relations based on semantic relatedness among examples, and explicitly steer adversarial representations out of the semantic neighborhood of the original event and toward dissimilar ones within each modality. As event semantics are shared across heterogeneous audio-visual tasks, such semantic shifts can mislead diverse downstream models without relying on any task objective.

Specifically, for each segment s and modality m, we pre-construct semantically similar references $\mathcal{R}^{\text{sim}}_{s,m}$ and semantically dissimilar references $\mathcal{R}^{\text{dis}}_{s,m}$ using clean features extracted by the surrogate encoder ψ_m from all segments of all videos in the test set $\mathcal{D}_{\text{test}}$. Taking AVEL as an example, reference features from segments sharing the query segment's event label are assigned to $\mathcal{R}^{\text{sim}}_{s,m}$, whereas those from segments with different event labels are assigned to $\mathcal{R}^{\text{dis}}_{s,m}$. Semantic relatedness can alternatively be estimated from neighborhoods in the clean feature space (Dwibedi et al., 2021). The reference features and their assignments remain fixed during attack generation.

For each segment and modality, we minimize

$$\mathcal{L}^{\text{intra}}_{s,m} = -\log \frac{\sum_{r \in \mathcal{R}^{\text{dis}}_{s,m}} \exp\big(\cos(g^{\text{adv}}_{s,m}, r)/\tau\big)}{\sum_{r \in \mathcal{R}^{\text{sim}}_{s,m} \cup \mathcal{R}^{\text{dis}}_{s,m}} \exp\big(\cos(g^{\text{adv}}_{s,m}, r)/\tau\big)} \tag{2}$$

Here, $\tau > 0$ is the temperature. Minimizing $\mathcal{L}^{\text{intra}}_{s,m}$ encourages $g^{\text{adv}}_{s,m}$ to align with semantically dissimilar references rather than its original semantic neighborhood.

Notably, contrastive learning has also been used to craft adversarial video examples. TVA (Lu et al., 2026) constructs contrastive pairs based on clean–adversarial frame correspondence, although non-corresponding frames may share the same semantics. In contrast, our method defines contrastive relations based on semantic relatedness within each modality. Moreover, while TVA perturbs only the visual input, our attack jointly perturbs audio and video and disrupts their cross-modal alignment.

3.3 Inter-Modal Misalignment Attack

The intra-modal attack perturbs each modality separately and leaves the correspondence between synchronized audio and visual streams unconstrained. However, audio-visual models rely on cross-modal agreement, where one modality can corroborate or compensate for the other, and mismatched audio-visual semantics can lead to erroneous predictions (Lee et al., 2022; Ye et al., 2026). Moreover, lower audio-visual feature similarity has been associated with stronger attack transferability (Zhang et al., 2025). We therefore explicitly break the correspondence between synchronized audio and visual features, turning complementary cues into conflicting evidence for cross-modal fusion.

Since the surrogate features are not necessarily aligned and may differ in dimensionality, we learn modality-specific projectors $\text{proj}_m : \mathbb{R}^{D_m} \to \mathbb{R}^d$, where D_m is the output dimension of encoder ψ_m and d is the shared projection dimension. Specifically, before attack generation, we train the two projectors on $\mathcal{D}_{\text{aux}}$, adopting the segment-level pairing strategy (Iashin et al., 2024). Inputs from the same interval of the same video form a positive pair, while all nonmatching segment pairs in the batch serve as negatives, including those from different intervals of the same video. Writing $u_{b,m} = \text{proj}_m(g_{b,m})$, we minimize the average of audio-to-visual and visual-to-audio InfoNCE losses with logits $u^\top_{b,a} u_{b',v}/\tau_{\text{p}}$, where b, b' index auxiliary batch segments and $\tau_{\text{p}} > 0$ is the temperature.

During the attack, with $u_{s,m}^{\mathrm{adv}} = \mathrm{proj}_m(g_{s,m}^{\mathrm{adv}})$, we minimize the segment-wise inter-modal loss

$$\mathcal{L}_s^{\mathrm{inter}} = \left(u_{s,a}^{\mathrm{adv}}\right)^\top u_{s,v}^{\mathrm{adv}} \tag{3}$$

This objective reduces the similarity between synchronized modalities in the learned space, complementing the intra-modal semantic shifts.

3.4 Joint Optimization

For each segment s and active modality $m \in \{a, v\}$, we then combine the intra-modal and inter-modal objectives as

$$\mathcal{L}_{s,m}^{\mathrm{attack}} = \mathcal{L}_{s,m}^{\mathrm{intra}} + \lambda \cdot \mathcal{L}_s^{\mathrm{inter}} \tag{4}$$

where $\lambda \geq 0$ balances the two objectives. Within each segment, we alternate between audio and visual perturbation updates. At each update, the other modality is held fixed, and gradients propagate through the frozen encoders and projections to the active input perturbation:

$$\delta_{s,m} \leftarrow \mathrm{clip}_{s,m}\left(\delta_{s,m} - \eta_m \,\mathrm{sign}\left(\nabla_{\delta_{s,m}} \mathcal{L}_{s,m}^{\mathrm{attack}}\right)\right) \tag{5}$$

Here, $\eta_m > 0$ is the modality-specific step size, and $\mathrm{clip}_{s,m}$ enforces the perturbation budget and valid input ranges. Each modality update uses the latest perturbations of both modalities.

In practice, we jointly optimize all segment-specific perturbations $\boldsymbol{\delta}_m = (\delta_{s,m})_{s=1}^S$ within each video by minimizing $\mathcal{L}_m^{\mathrm{attack}} = \frac{1}{S}\sum_{s=1}^S \mathcal{L}_{s,m}^{\mathrm{attack}}$ with the same alternating PGD updates. This enables batched segment processing for computational efficiency without changing the segment-wise objectives or perturbation parameterization.

After optimization, we add $\delta_{s,a}$ to the audio waveform of each segment s and apply the same visual perturbation $\delta_{s,v}$ to all T_s frames within that segment, yielding $v_{s,t}^{\mathrm{adv}} = \mathrm{clip}_v(v_{s,t} + \delta_{s,v})$ for $t = 1, \cdots, T_s$. Here, $v_{s,t}$ denotes the t-th frame of segment s (Huang et al., 2026), and clip_v enforces valid pixel ranges. The perturbed audio and video segments are then reassembled in their original temporal order to obtain the complete adversarial audio-visual input.

Discussion. Existing multimodal attacks mainly target vision-language models (Lu et al., 2023; Wang et al., 2024), whose encoders are often pretrained into a shared embedding space with sample-level image–text correspondence. In contrast, audio-visual models typically rely on independently pretrained audio and visual encoders without such a space, and their correspondence is defined by temporal synchronization. AV-RIFT therefore learns lightweight projectors from auxiliary paired data to bridge the two encoders, and disrupts cross-modal correspondence at the segment level, making it tailored to audio-visual data rather than a direct extension of vision-language attacks.

4 Experiment

4.1 Experimental Setup

Datasets and Metrics. We evaluate 5 representative audio-visual tasks: event localization on the AVE dataset (Tian et al., 2018), segmentation on the single source subset of AVSBench (Zhou et al., 2022b), classification on the Kinetics-Sounds dataset (Arandjelovic & Zisserman, 2017), retrieval on the VEGAS dataset (Zhou et al., 2018), and question answering on the MUSIC-AVQA dataset (Li et al., 2022a). We report accuracy for AVEL, AVER, and AVQA, mIoU for AVS, and bidirectional mAP for AVR. Attack success rate (ASR) is computed according to the task-specific output structure of each task. See the appendix for details.

Evaluated Models. Our benchmark contains 16 task-specific models. For AVEL, we evaluate A+V-att (Tian et al., 2018), PSP (Zhou et al., 2022a), CMBS (Xia & Zhao, 2022), and CACE-Net (He et al., 2024). For AVER, following Tian & Xu (2021), we pair a ResNet-18 visual backbone with an ANet audio backbone (R + A) and consider three fusion strategies, including sum, concatenation, and FiLM (Perez et al., 2018). For AVS, we include TPAVI (Zhou et al., 2022b), UFE-TPAVI (Liu et al., 2024), AVSegFormer (Gao et al., 2024), and CAVP (Chen et al., 2024). For AVR, we evaluate AML (Zheng et al., 2021), CMR (Zeng et al., 2020), and AVID-CMA (Morgado et al., 2021). For

Table 1: Black-box transfer attack results (%) on 16 **task-specific** audio-visual models across 5 tasks. I2V (Wei et al., 2022b) and SoundBreak (SB) (Hussain et al., 2026) attack the visual and audio modalities, respectively, while I2V + SB perturbs both. Perf. denotes accuracy (AVEL/AVER/AVQA), mIoU (AVS), or mAP (AVR); ASR denotes the attack success rate. The best results among attacks in each Avg. row are shown in **bold**.

Task	Target model	Clean	I2V		SoundBreak		I2V + SB		**AV-RIFT (Ours)**	
			Perf.↓	ASR↑	Perf.↓	ASR↑	Perf.↓	ASR↑	Perf.↓	ASR↑
AVEL	A+V-att (Tian et al., 2018)	66.99	64.78	8.39	21.04	77.39	19.68	79.13	5.20	95.66
	PSP (Zhou et al., 2022a)	72.51	68.68	12.59	46.09	43.12	35.12	61.68	11.17	89.85
	CMBS (Xia & Zhao, 2022)	74.45	70.17	10.66	46.84	45.04	33.58	61.71	5.45	94.29
	CACE-Net (He et al., 2024)	77.51	73.33	10.40	50.60	41.91	36.82	59.31	7.89	92.84
	Avg.	72.87	69.24	10.51	41.14	51.87	31.30	65.46	**7.43**	**93.16**
AVS	TPAVI (Zhou et al., 2022b)	71.54	58.36	9.11	49.96	16.24	44.18	24.32	42.50	25.68
	AVSegFormer (Gao et al., 2024)	78.96	72.34	3.65	75.62	3.43	49.43	20.54	51.60	18.76
	UFE-TPAVI (Liu et al., 2024)	78.15	72.51	5.11	71.27	6.49	54.16	17.81	49.33	23.46
	CAVP (Chen et al., 2024)	78.78	70.06	5.51	78.56	2.92	63.63	9.73	56.66	13.97
	Avg.	76.86	68.32	5.85	68.85	7.27	52.85	18.10	**50.02**	**20.47**
AVER	R + A + Concat (Tian & Xu, 2021)	72.49	62.82	18.89	69.43	8.36	47.26	39.32	40.30	47.71
	R + A + Sum (Tian & Xu, 2021)	72.33	63.11	18.71	69.01	8.11	49.71	37.24	42.01	46.19
	R + A + FiLM (Perez et al., 2018)	72.52	61.05	22.03	68.11	9.06	46.78	41.16	38.82	50.78
	Avg.	72.45	62.33	19.88	68.85	8.51	47.92	39.24	**40.38**	**48.23**
AVR	AML (Zheng et al., 2021)	73.31	64.69	18.04	23.42	65.84	47.10	49.27	31.71	65.95
	CMR (Zeng et al., 2020)	74.14	63.73	23.15	21.84	70.14	45.69	51.16	47.44	45.77
	AVID-CMA (Morgado et al., 2021)	85.60	75.13	13.34	64.44	21.83	65.65	28.48	29.91	68.94
	Avg.	77.68	67.85	18.18	36.57	52.60	52.81	42.97	**36.35**	**60.63**
AVQA	STG (Li et al., 2022a)	68.61	65.23	7.21	49.69	33.16	46.95	38.60	32.42	62.36
	QA-TIGER (Kim et al., 2025)	76.60	77.28	3.37	70.70	14.74	70.57	15.01	64.46	25.81
	Avg.	72.61	71.26	5.29	60.20	23.95	58.76	26.81	**48.44**	**44.08**

AVQA, we consider STG (Li et al., 2022a) and QA-TIGER (Kim et al., 2025). In addition, to examine transfer beyond task-specific architectures, we evaluate DG-SCT (Duan et al., 2023), Crab (Du et al., 2025), WAVE-7B (Tang et al., 2026), and PE-AV (Vyas et al., 2026) on their supported tasks.

Attack Baselines. For a fair comparison, we include I2V (Wei et al., 2022b) and SoundBreak (Hussain et al., 2026), which perform feature attacks on the visual and audio modalities, respectively, as well as their combined variant, I2V + SoundBreak. On AVER, we further compare with the unimodal PGD (Madry et al., 2017) and MI-FGSM (Dong et al., 2018) attacks and the audio-visual TMA (Zhang et al., 2025) attack.

Implementation Details. AV-RIFT operates on one-second synchronized audio-visual segments with frozen ResNet-50 and VGGish surrogate encoders. Following Iashin et al. (2024), we pretrain the AV-RIFT projectors on paired VGGSound (Chen et al., 2020) clips using modality-specific MLP heads and symmetric InfoNCE with $\tau_{\mathrm{p}} = 0.07$. The projectors are optimized with AdamW (learning rate 10^{-3}, weight decay 10^{-4}). We use 50-step alternating PGD with random initialization, perturbation budgets $(\epsilon_v, \epsilon_a) = (16/255, 0.04)$, step sizes $(1/255, 0.002)$, and loss balance $\lambda = 1$. The same budgets are used for all attack baselines. During transfer attacks, the surrogate encoders, projectors, and clean reference features remain fixed. Other baselines follow their original settings.

4.2 Attack Results

Attacks on Task-specific Models. As shown in Table 1, AV-RIFT achieves the highest average ASR on all five tasks, outperforming the strongest baseline by 12.87% on average. On AVEL, it reduces accuracy from 72.87% to 7.43%, yielding an ASR of 93.16%. In contrast, the baselines are strongly modality-dependent: I2V is more effective on AVER, SoundBreak on AVEL, AVR, and AVQA, and I2V + SB even underperforms SoundBreak alone on AVR. By jointly exploiting intra-modal semantics and cross-modal correspondence, AV-RIFT avoids this dependence, and its advantage also holds with ViT-S and Whisper as surrogates (Appendix A.1).

Table 2: Attack results (%) on 4 **general** audio-visual models across their supported tasks under black-box and gray-box settings. The black-box setting uses CLIP and BEATs as surrogate encoders, while the gray-box setting uses the target model's own visual and audio encoders.

Attack setting	Attack	Metric	DG-SCT (Duan et al., 2023)			Crab (Du et al., 2025)			WAVE (Tang et al., 2026)				PE-AV (Vyas et al., 2026)	
			AVEL	AVS	AVQA	AVEL	AVS	AVQA	AVEL	AVER	AVR	AVQA	AVER	AVR
–	Clean	Perf.	68.26	80.89	73.24	92.04	63.76	81.86	72.49	82.25	78.66	80.97	89.01	62.07
Black-box	I2V + SB	Perf.↓	67.36	76.35	71.80	72.64	54.51	79.74	67.84	74.71	43.44	79.24	85.41	43.33
		ASR↑	6.77	3.03	3.51	22.70	8.73	8.32	19.17	14.86	44.69	2.28	6.33	41.78
	AV-RIFT (Ours)	Perf.↓	**61.04**	**73.29**	**63.18**	**59.20**	**51.61**	**74.26**	**60.30**	**70.36**	**16.25**	**75.30**	**82.47**	**34.46**
		ASR↑	**17.18**	**8.41**	**17.84**	**37.30**	**8.78**	**15.31**	**37.82**	**18.86**	**88.97**	**19.68**	**9.37**	**58.69**
Gray-box	I2V + SB	Perf.↓	66.94	75.75	70.30	66.17	52.24	76.62	65.82	74.21	38.36	76.50	80.58	38.03
		ASR↑	9.90	3.92	5.91	29.73	10.73	13.78	25.09	15.90	51.01	10.74	12.71	52.91
	AV-RIFT (Ours)	Perf.↓	**59.83**	**73.65**	**60.18**	**42.78**	**52.08**	**66.03**	**56.19**	**71.08**	**13.05**	**73.19**	**77.25**	**20.90**
		ASR↑	**25.04**	**5.86**	**22.84**	**48.97**	**12.54**	**38.02**	**45.62**	**20.71**	**94.08**	**23.85**	**18.10**	**76.86**

Table 3: Robustness of attacks against input-transformation defenses on **task-specific** models. Results (%) are averaged over the target models of each task. Undefended denotes no defense. BDR + WQ pairs visual bit-depth reduction with audio waveform quantization, and JPEG + MP3 pairs JPEG compression with MP3 compression. Lower performance indicates a stronger attack. Bold marks the stronger attack under each setting.

Attack	AVEL			AVS			AVR			AVQA		
	Undefended	BDR + WQ	JPEG + MP3	Undefended	BDR + WQ	JPEG + MP3	Undefended	BDR + WQ	JPEG + MP3	Undefended	BDR + WQ	JPEG + MP3
Clean	72.87	71.03	72.09	76.86	74.39	73.88	77.68	74.82	76.53	72.61	71.22	71.59
I2V + SB	31.30	31.65	41.01	52.85	54.32	62.70	52.81	52.95	55.81	58.76	59.07	64.54
AV-RIFT (Ours)	**7.43**	**7.62**	**11.22**	**50.02**	**52.54**	**45.25**	**38.01**	**38.87**	**41.80**	**48.44**	**48.70**	**50.45**

Attacks on General Audio-Visual Models. We consider both black-box and gray-box (Liu et al., 2026) settings. The black-box setting uses CLIP and BEATs as surrogate visual and audio encoders, respectively, while the gray-box setting directly uses the target model's own visual and audio encoders to construct the attack objectives. As shown in Table 2, AV-RIFT consistently outperforms I2V + SB under both settings, improving ASR by 13.00% and 15.85% on average, respectively. The attack remains highly effective on large-model-based systems such as Crab and WAVE; for example, on WAVE for AVR, AV-RIFT reaches 88.97% and 94.08% ASR under the black-box and gray-box settings, compared with 44.69% and 51.01% for I2V + SB. Access to the target encoders further strengthens AV-RIFT in 11 of the 12 cases, e.g., raising its ASR on Crab for AVQA from 15.31% to 38.02%. These results indicate that the semantic vulnerability exploited by AV-RIFT extends beyond task-specific architectures to general audio-visual models.

Comparison with Existing Audio-Visual Attacks. We further compare AV-RIFT with PGD (Madry et al., 2017), MI-FGSM (Dong et al., 2018), and the audio-visual attack TMA (Zhang et al., 2025) on AVER, the task for which TMA is defined. Following the task-aware setting of TMA, all attacks have white-box access to R + A + Concat and optimize its classification loss, and the resulting perturbations are transferred to R + A + Sum and R + A + FiLM. For a fair comparison, we add the same classification loss to the AV-RIFT objective. As shown in Table 4, PGD and MI-FGSM nearly succeed on the source model but transfer poorly, with ASR dropping from above 90% to 55–64%. TMA transfers better by additionally disrupting temporal and cross-modal consistency, reaching 88.26% and 93.03% ASR on the two transfer targets. AV-RIFT achieves the highest ASR on all three models, with 98.22% on the source model and 93.11% and 96.16% on the transfer targets. This shows that the intra-modal semantic and inter-modal misalignment objectives complement the task loss and improve transferability.

Table 4: AVER results (%) under the TMA setting. All attacks use the Concat classification loss with white-box access (†) and transfer to Sum and FiLM.

Attack	Concat Acc.↓/ASR↑	Sum Acc.↓/ASR↑	FiLM Acc.↓/ASR↑
Clean	72.49/–	72.33/–	72.52/–
PGD	6.67/90.63†	61.63/55.16	56.83/60.49
MI-FGSM	3.58/93.98†	60.80/58.04	54.90/64.08
TMA	2.79/97.05†	7.65/88.26	5.28/93.03
AV-RIFT (Ours)	**2.24/98.22**†	**4.01/93.11**	**3.61/96.16**

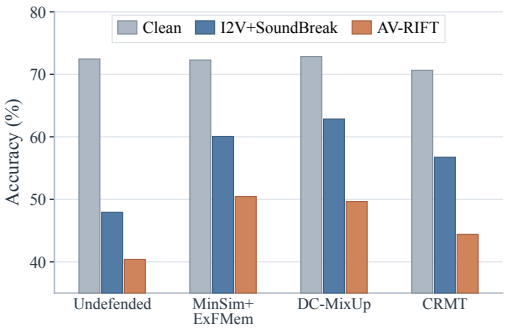


Figure 3: **Attack robustness against audio-visual defenses on AVER.** Accuracy (%) is averaged over the three R + A models. Lower accuracy indicates a stronger attack.

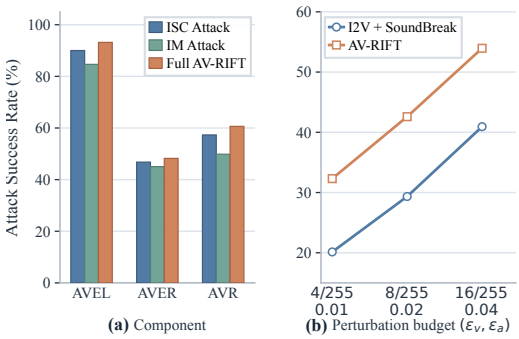


Figure 4: **Ablation study of AV-RIFT.** (a) ASR (%) with the ISC attack, the IM attack, or both. (b) Average ASR (%) under increasing budgets (ϵ_v, ϵ_a). Higher ASR indicates a stronger attack.

4.3 Defense Results

Robustness against Input-Transformation Defenses. We further evaluate our attack against two input-transformation defenses across tasks: visual bit-depth reduction (Xu et al., 2017) paired with audio waveform quantization (Hussain et al., 2021) (BDR + WQ), and JPEG compression (Dziugaite et al., 2016) paired with audio MP3 compression (Chen et al., 2022) (JPEG + MP3). As shown in Table 3, AV-RIFT retains higher attack effectiveness than I2V + SoundBreak, reducing the average defended performance by 11.96 and 20.42 percentage points under BDR + WQ and JPEG + MP3, respectively. These results demonstrate its robustness to common input transformations.

Robustness against Audio-Visual Defenses. We evaluate our attack on AVER against three audio-visual defenses: MinSim + ExFMem (Tian & Xu, 2021), CRMT (Yang et al., 2024), and DC-MixUp (Li et al., 2022b). Figure 3 compares these three defenses with the undefended model. Clean-input accuracy remains stable across methods. Under I2V + SoundBreak, existing defenses substantially recover performance, whereas under AV-RIFT, their gains are much more limited over the undefended model, with accuracy remaining below 51%. This gap indicates that AV-RIFT generates adversarial examples that are more resistant to existing defenses.

4.4 Ablation Study

Figure 4(a) ablates the two components of AV-RIFT: the intra-modal semantics-aware contrastive (ISC) attack and the inter-modal misalignment (IM) attack. Using either one alone consistently underperforms the full AV-RIFT, which reaches ASRs of 93.16%, 48.23%, and 60.63% on AVEL, AVER, and AVR, with average gains of 2.64% over the ISC attack and 7.49% over the IM attack. This indicates that the two attacks are complementary, with ISC being the stronger one.

Figure 4(b) further varies the perturbation budget (ϵ_v, ϵ_a) from $(4/255, 0.01)$ to $(16/255, 0.04)$. While both attacks strengthen with larger budgets, AV-RIFT consistently outperforms I2V + SoundBreak by 12.81% on average, remaining effective under both constrained and larger budgets.

5 Conclusion

In this work, we study a practical threat to audio-visual models, where the attacker cannot access downstream models and relies only on public pretrained encoders as surrogates. We propose **AV-RIFT**, which moves beyond instance-level feature disruption by exploiting semantic relationships within each modality and the correspondence between synchronized modalities. Extensive experiments across multiple tasks and diverse model architectures, including audio-visual large language models, show that AV-RIFT consistently outperforms existing attacks and remains effective under existing defenses. These results reveal the vulnerability of audio-visual models to cross-task attacks, which deserves greater attention.

AI USE STATEMENT

We used generative AI tools in two ways. First, at the early stage of this project, we used them to help search for related papers, mainly when preparing the related work on audio-visual learning, audio-visual adversarial attacks, and transfer attacks on pretrained encoders. Every suggested paper was read and verified by the authors before being cited. Second, we used them to polish the language of the manuscript, such as correcting grammar, refining wording, and improving the flow between sentences. The research ideas, method design, experiments, and analysis were carried out by the authors. The authors have reviewed all AI-assisted content and take full responsibility for the content of this paper.

REFERENCES

Sajjad Abdoli, Luiz G Hafemann, Jerome Rony, Ismail Ben Ayed, Patrick Cardinal, and Alessandro L Koerich. Universal adversarial audio perturbations. *arXiv preprint arXiv:1908.03173*, 2019.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617. IEEE, 2017.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, Feng Wang, and Jiashui Wang. Towards understanding and mitigating audio adversarial examples for speaker recognition. *IEEE Transactions on Dependable and Secure Computing*, 20(5):3970–3987, 2022.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26487–26497. IEEE, 2024.

Marco Cristani, Manuele Bicego, and Vittorio Murino. Audio-visual event recognition in surveillance video sequences. *IEEE transactions on multimedia*, 9(2):257–267, 2007.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193. IEEE, 2018.

Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, and Di Hu. Crab: A unified audio-visual scene understanding model with explicit cooperation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18804–18814. IEEE, 2025.

Haoyi Duan, Yan Xia, Zhou Mingze, Li Tang, Jieming Zhu, and Zhou Zhao. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. *Advances in Neural Information Processing Systems*, 36:56075–56094, 2023.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9568–9577. IEEE, 2021.

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12155–12163, 2024.

Xiang He, Xiangxi Liu, Yang Li, Dongcheng Zhao, Guobin Shen, Qingqun Kong, Xin Yang, and Yi Zeng. Cace-net: Co-guidance attention and contrastive enhancement for effective audio-visual event localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 985–993, 2024.

Linhao Huang, Xue Jiang, Zhiqiang Wang, Wentao Mo, Xi Xiao, Yong-Jie Yin, Bo Han, and Feng Zheng. Transferability of adversarial attacks in video-based mllms: A cross-modal image-to-video approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5067–5075, 2026.

Aafiya Shamshad Hussain, Gaurav Srivastava, Alvi Md Ishmam, Zaber Ibn Abdul Hakim, and Chris Thomas. Soundbreak: A systematic study of audio-only adversarial attacks on trimodal models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27635–27663, 2026.

Shehzeen Hussain, Paarth Neekhara, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. {WaveGuard}: Understanding and mitigating audio adversarial examples. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2273–2290, 2021.

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

Hongyeob Kim, Inyoung Jung, Dayoon Suh, Youjia Zhang, Sangmin Lee, and Sungeun Hong. Question-aware gaussian experts for audio-visual question answering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13681–13690. IEEE, 2025.

Sangmin Lee, Sungjune Park, and Yong Man Ro. Audio-visual mismatch-aware video retrieval via association and adjustment. In *European Conference on Computer Vision*, pp. 497–514. Springer, 2022.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19086–19096. IEEE, 2022a.

Juncheng B Li, Shuhui Qu, Xinjian Li, Po-Yao Bernie Huang, and Florian Metze. On adversarial robustness of large-scale audio visual learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 231–235. IEEE, 2022b.

Daizong Liu, Xiaowen Cai, Junhao Dong, Zhongliang Guo, Xiaoye Qu, Runwei Guan, Xiang Fang, and Dengpan Ye. Attacking gray-box large vision-language models with adaptive svd-structured adversarial alignment. In *Forty-third International Conference on Machine Learning*, 2026.

Jinxiang Liu, Yikun Liu, Fei Zhang, Chen Ju, Ya Zhang, and Yanfeng Wang. Audio-visual segmentation via unlabeled frame exploitation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26318–26329. IEEE, 2024.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 102–111. IEEE, 2023.

Hui Lu, Yi Yu, Song Xia, Yiming Yang, Deepu Rajan, Boon Poh Ng, Alex Kot, and Xudong Jiang. From pretrain to pain: Adversarial vulnerability of video foundation models without task knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7548–7556, 2026.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12470–12481. IEEE, 2021.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pp. 5231–5240. PMLR, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Changli Tang, Qinfan Xiao, Ke Mei, Tianyi Wang, Fengyun Rao, and Chao Zhang. Wave: learning unified & versatile audio-visual embeddings with multimodal llm. In *International Conference on Learning Representations*, volume 2026, pp. 61596–61612, 2026.

Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5597–5607. IEEE, 2021.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European conference on computer vision*, pp. 252–268. Springer, 2018.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matthew Le, Piotr Dollár, Christoph Feichtenhofer, et al. Pushing the frontier of audio-visual perception with large-scale multimodal correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30172–30182, 2026.

Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1722–1740. IEEE, 2024.

Xiaosen Wang, Zhijin Ge, Bohan Liu, Zheng Fang, Fengfan Zhou, Ruixuan Zhang, Shaokang Wang, and Yuyang Luo. Devling into Adversarial Transferability on Image Classification: Review, Benchmark, and Evaluation. *arXiv preprint arXiv:2602.23117*, 2026.

Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022a.

Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Cross-modal transferable adversarial attacks from images to videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15044–15053. IEEE, 2022b.

Jinming Wen, Xinyi Wu, Shuai Zhao, Yanhao Jia, and Yuwen Li. Investigating vulnerabilities and defenses against audio-visual attacks: A comprehensive survey emphasizing multimodal models. *arXiv preprint arXiv:2506.11521*, 2025.

Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19957–19966. IEEE, 2022.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2725–2734. IEEE, 2019.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3339–3348. IEEE, 2021.

Zequn Yang, Yake Wei, Ce Liang, and Di Hu. Quantifying and enhancing multi-modal robustness with modality preference. In *International Conference on Learning Representations*, volume 2024, pp. 15860–15872, 2024.

Qilang Ye, Zitong Yu, Rui Shao, Yawen Cui, Xiangui Kang, Xin Liu, Philip Torr, and Xiaochun Cao. Cat+: Investigating and enhancing audio-visual understanding in large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8674–8690, 2025.

Qilang Ye, Wei Zeng, Meng Liu, Jie Zhang, Yupeng Hu, Zitong Yu, and Yu Zhou. When eyes and ears disagree: Can mllms discern audio-visual confusion? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11955–11963, 2026.

Donghuo Zeng, Yi Yu, and Keizo Oyama. Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(3):1–23, 2020.

Zeliang Zhang, Susan Liang, Daiki Shimada, and Chenliang Xu. Rethinking audio-visual adversarial vulnerability from temporal and modality perspectives. In *International Conference on Learning Representations*, volume 2025, pp. 20199–20215, 2025.

Aihua Zheng, Menglan Hu, Bo Jiang, Yan Huang, Yan Yan, and Bin Luo. Adversarial-metric learning for audio-visual cross-modal matching. *IEEE Transactions on Multimedia*, 24:338–351, 2021.

Jinxing Zhou, Dan Guo, and Meng Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7239–7257, 2022a.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2022b.

Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3550–3558. IEEE, 2018.

Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18(3):351–376, 2021.

Appendix overview. The appendix is organized to present the most direct evidence first, followed by the details needed to reproduce and interpret our results.

- **Appendix A (Additional Experimental Results)** further demonstrates the effectiveness and generality of AV-RIFT. It shows that AV-RIFT remains superior with alternative surrogate encoders and visualizes the adversarial examples it generates across tasks.
- **Appendix B (Attack Details of AV-RIFT)** provides the complete attack procedure, including preprocessing, semantic reference construction, projector training, perturbation optimization, and the algorithm summary.
- **Appendix C (Experimental Details)** specifies the evaluation conventions, the implementation of baseline attacks, and the task-specific definitions of performance and ASR, which ensure a fair comparison across heterogeneous tasks.

A Additional Experimental Results

This section provides additional attack results that complement Section 4.1: transferability with alternative surrogate encoders (Appendix A.1) and visualizations of the generated adversarial examples (Appendix A.2).

A.1 TRANSFERABILITY WITH ALTERNATIVE SURROGATE ENCODERS

Table 5: Transfer attack results (%) on **task-specific** models with alternative surrogate encoders, i.e., ViT-S for video and Whisper for audio, instead of the default ResNet-50 and VGGish used in Table 1. Perf. denotes accuracy (AVEL/AVER), mIoU (AVS), or mAP (AVR); ASR follows the task-specific definition in Table 1. The average results of AV-RIFT are shown in **bold**.

Task	Target model	Clean	I2V + SB		**AV-RIFT (Ours)**	
			Perf.↓	ASR↑	Perf.↓	ASR↑
AVEL	A+V-att (Tian et al., 2018)	66.99	43.36	44.56	29.70	62.38
	PSP (Zhou et al., 2022a)	72.51	59.00	27.68	42.06	48.95
	CMBS (Xia & Zhao, 2022)	74.45	60.75	25.76	44.35	45.44
	CACE-Net (He et al., 2024)	77.51	63.61	24.04	46.59	43.58
	Avg.	72.87	56.68	30.51	**40.68**	**50.09**
AVS	TPAVI (Zhou et al., 2022b)	71.54	63.43	8.73	58.91	12.38
	AVSegFormer (Gao et al., 2024)	78.96	69.64	4.92	66.47	6.84
	UFE-TPAVI (Liu et al., 2024)	78.15	70.11	5.95	65.27	9.30
	CAVP (Chen et al., 2024)	78.78	67.39	7.27	62.71	9.57
	Avg.	76.86	67.64	6.72	**63.34**	**9.52**
AVER	R + A + Concat (Tian & Xu, 2021)	72.49	52.64	31.54	47.29	37.83
	R + A + Sum (Tian & Xu, 2021)	72.33	55.41	28.65	50.03	35.07
	R + A + FiLM (Perez et al., 2018)	72.52	50.26	35.39	44.91	41.94
	Avg.	72.45	52.77	31.86	**47.41**	**38.28**
AVR	AML (Zheng et al., 2021)	73.31	58.01	34.18	44.48	54.03
	CMR (Zeng et al., 2020)	74.14	57.25	35.60	42.55	57.01
	AVID-CMA (Morgado et al., 2021)	85.60	66.64	27.80	52.54	43.58
	Avg.	77.68	60.63	32.53	**46.52**	**51.54**

To verify that the advantage of AV-RIFT does not depend on a particular surrogate, we repeat the transfer attack on task-specific models with a different pair of public encoders. The main results in Table 1 use ResNet-50 and VGGish as surrogate encoders. Since the threat model in Section 3.1 does not require the surrogates to match the encoders of the downstream models, an attacker may choose any publicly available encoder pair. Here, we replace the surrogates with a ViT-S visual encoder and a Whisper audio encoder. Both are Transformer-based and thus differ in architecture from the CNN-based default pair, and Whisper is trained primarily for speech recognition rather than for the sound events targeted by these tasks. For this encoder pair, we retrain the projectors on the same auxiliary data following Appendix B.3 and re-extract the clean reference features with the new surrogates. All other settings, including the perturbation budgets, step sizes, number of iterations, and target models, are identical to those in Section 4.1. We compare against I2V + SB implemented with the same surrogate encoders.

Table 5 reports the results. AV-RIFT remains consistently stronger than I2V + SB under the alternative surrogates, and the advantage holds for every individual target model. On AVEL, AV-RIFT reduces the average accuracy to 40.68%, compared with 56.68% for I2V + SB, and improves the average ASR from 30.51% to 50.09%. On AVS, it reduces the average mIoU from 67.64% to 63.34% and improves the average ASR from 6.72% to 9.52%. On AVER, it reduces the average accuracy to 47.41%, compared with 52.77% for I2V + SB, and improves the average ASR from 31.86% to 38.28%. On AVR, it reduces the average mAP from 60.63% to 46.52% and improves the average ASR from 32.53% to 51.54%.

Compared with Table 1, both attacks become weaker with the alternative surrogates, e.g., the average AVEL ASR of AV-RIFT decreases from 93.16% to 50.09%. We attribute this to the larger gap between the surrogate and target representations. Many target models build on CNN visual backbones and VGGish-style audio features, which are closer to the default surrogates. Nevertheless,

AV-RIFT preserves a clear margin over I2V + SB, e.g., 19.58 percentage points in average ASR on AVEL. This indicates that its advantage stems from the attack objectives, i.e., semantics-aware intra-modal disruption and inter-modal misalignment, rather than from a specific surrogate choice. It further supports the practical threat considered in this work: an attacker can mount AV-RIFT with different publicly available encoders, without knowing which encoders the downstream models use.

A.2 Visualization of Adversarial Examples

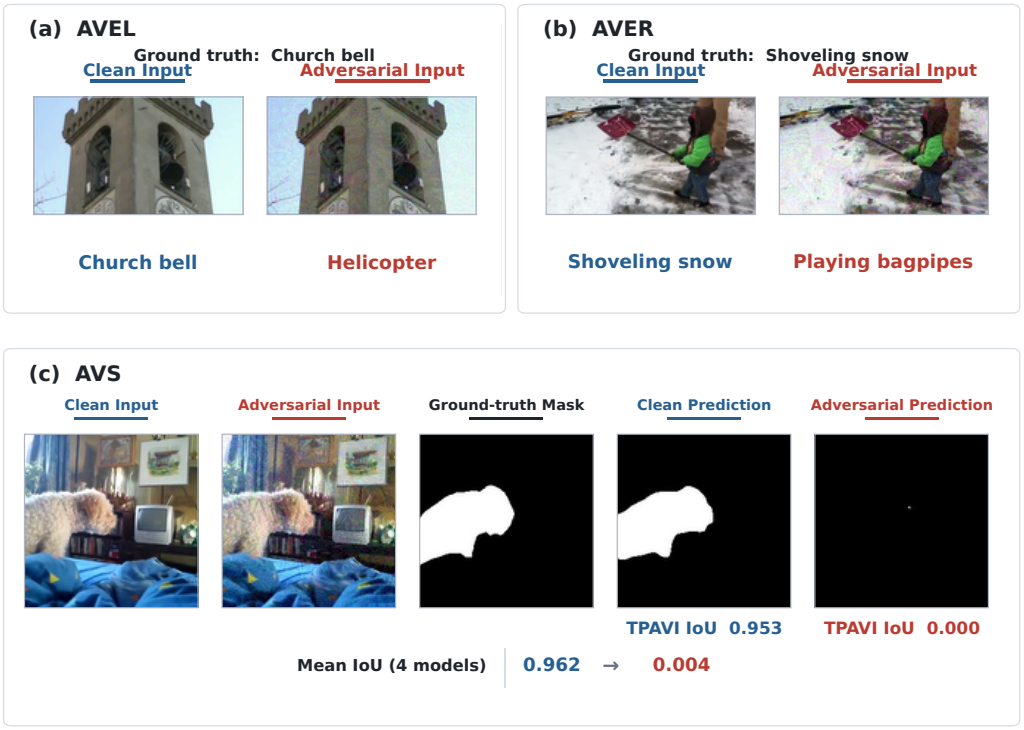


Figure 5: **Adversarial examples generated by AV-RIFT on three audio-visual tasks.** (a) AVEL and (b) AVER show the clean and adversarial frames with the corresponding predicted event labels. (c) AVS shows the clean and adversarial frames, the ground-truth mask, and the TPAVI predictions with their IoU; the mean IoU over the four AVS models is reported below. The audio stream is also perturbed but not shown.

Figure 5 visualizes adversarial examples generated by AV-RIFT and their effect on target models of three tasks. The perturbations are crafted solely on the surrogate encoders with $(\epsilon_v, \epsilon_a) = (16/255, 0.04)$, without access to any target model. Although the scene content remains clearly recognizable in the adversarial frames, the target predictions change substantially. On AVEL, a church-bell segment is localized as *helicopter*; on AVER, a clip of shoveling snow is classified as *playing bagpipes*. On AVS, the sounding-object mask predicted by TPAVI almost vanishes, with its IoU dropping from 0.953 to 0.000, and the mean IoU over the four AVS models drops from 0.962 to 0.004. The same attack procedure thus disrupts event localization, classification, and pixel-level segmentation, illustrating the task-agnostic transferability of AV-RIFT.

B Attack Details of AV-RIFT

The following details specify the attack protocol used for the transfer results in the main paper. Unless stated otherwise, all attack components use the same frozen surrogate encoders, segment partition, perturbation budgets, and optimization schedule; only the target models used for evaluation change.

B.1 Attack Setting and Common Preprocessing

The adversary does not query, differentiate through, or otherwise use any downstream target model. It has access to the labeled evaluation split used to construct the clean reference bank, the frozen surrogate encoders, and the auxiliary paired data used to train the projection heads. Reference features are extracted once before optimization and are detached from the computation graph. Thus, the attack uses labels only to define semantic reference membership and never uses target-model outputs or gradients.

We divide each audio-visual clip into contiguous, synchronized one-second segments. For each segment, the default preprocessing samples the center video frame and crops or pads the corresponding waveform to 16 kHz. Video inputs are normalized to $[0, 1]$ and audio waveforms to $[-1, 1]$. If a clip ends in a partial segment, its audio is padded only for feature extraction; padding is removed when the final waveform is exported. The surrogate feature extractor pools token, patch, or within-window time axes inside each segment, leaving one vector per segment. In the default ResNet-50/VGGish instantiation, these vectors have dimensions $D_v = 2048$ and $D_a = 128$, respectively.

The attack is evaluated on the five task families described in Section 4.1. Segment-level perturbations are optimized jointly within each input video, but the same visual perturbation is applied to every frame belonging to a segment when the adversarial video is reconstructed. The auxiliary projection-training split is kept disjoint from the downstream training and evaluation splits; its category labels and downstream annotations are not used by the projector objective.

B.2 Segment Representation and Semantic References

The same segment representation is used for clean references, auxiliary projection training, and adversarial inputs. For segment s and modality m, the clean and adversarial surrogate features in Section 3.1 are instantiated as $g_{s,m} = \mathrm{Pool}(\psi_m(x_{s,m}))$ and $g^{\mathrm{adv}}_{s,m} = \mathrm{Pool}(\psi_m(x^{\mathrm{adv}}_{s,m}))$, respectively; the pooling operator is omitted in the main text for brevity. The temporal index is retained throughout the attack, even though token or patch axes inside a segment are mean-pooled.

The sets $\mathcal{R}^{\mathrm{sim}}_{s,m}$ and $\mathcal{R}^{\mathrm{dis}}_{s,m}$ are built from clean features of the labeled evaluation split and encode semantic relatedness to the clean input rather than a target-model prediction. When a task has a single category or event label, references with the same label are assigned to $\mathcal{R}^{\mathrm{sim}}_{s,m}$ and references with a disjoint label to $\mathcal{R}^{\mathrm{dis}}_{s,m}$. For multi-label samples, any label overlap is treated as similar and disjoint nonempty label sets as dissimilar. The clean input's own features are included among the similar references, and the resulting sample-level assignments are shared by all of its segments.

For AVQA, where one video is associated with multiple questions, we aggregate the normalized answer, question type, question–answer group, and question template into a set of video-level semantic tags. Two videos are similar when their tag sets intersect and dissimilar when their nonempty tag sets are disjoint; the tag set is inherited by every segment of that video. If no usable annotation is available, we fall back to the clean feature-space neighborhood protocol (Dwibedi et al., 2021): same-source temporal segments form the similar pool and segments from other sources form the dissimilar pool, with the current segment removed from its same-source pool. This fallback is a proxy for semantic relatedness, not ground-truth equivalence.

The log-ratio in Eq. (2) uses every reference in the two selected sets rather than selecting a single target example. We compute it as the difference of two log-sum-exp terms for numerical stability and average the resulting segment losses within each video. Here, similar and dissimilar describe the relation to the clean input: the attack encourages similarity to the latter rather than adopting the positive-pair convention of representation learning. In all reported experiments, both reference sets are nonempty.

B.3 Projection Training on Auxiliary Paired Data

We learn the cross-modal projections from an auxiliary paired audio-visual dataset $\mathcal{D}_{\mathrm{aux}}$ that is separate from the downstream training and test sets. Temporal pairing supplies the supervision; category labels and downstream predictions are not used. For a fixed surrogate encoder pair, the same learned projections are used across all downstream tasks. Model selection uses held-out auxiliary data.

Each modality has a separate two-layer MLP consisting of input LayerNorm, a linear layer, GELU, and an output linear layer, followed by ℓ_2 normalization. For the ResNet-50/VGGish instantiation, the visual and audio input dimensions are 2048 and 128, respectively; both branches have hidden width 256 and output dimension 128. The encoders are frozen, and only the projection parameters are optimized with AdamW. For the reported checkpoints, AdamW uses learning rate 10^{-3} and weight decay 10^{-4}; training runs for at most 50 epochs with early stopping after 10 validation epochs without improvement. The checkpoint with the lowest held-out loss is retained.

Following Segment AVCLIP in Synchformer (Iashin et al., 2024), we contrast segment pairs rather than whole-video embeddings. Let $g_{b,m} = \mathrm{Pool}(\psi_m(x^{\mathrm{aux}}_{b,m})) \in \mathbb{R}^{D_m}$ denote the frozen pooled feature of auxiliary segment b, and let $u_{b,m} = \mathrm{proj}_m(g_{b,m}) \in \mathbb{R}^d$ denote its normalized projection. A minibatch with B videos and S_{aux} segments per video contains $Q = BS_{\mathrm{aux}}$ synchronized segment pairs after flattening the video and segment indices. Matching indices identify the same interval in the same video. The audio-to-video loss and symmetric training objective are

$$\mathcal{L}_{a\to v} = -\frac{1}{Q}\sum_{b=1}^{Q} \log \frac{\exp(u_{b,a}^\top u_{b,v}/\tau_{\mathrm{p}})}{\sum_{b'=1}^{Q}\exp(u_{b,a}^\top u_{b',v}/\tau_{\mathrm{p}})}, \qquad \mathcal{L}_{\mathrm{proj}} = \frac{\mathcal{L}_{a\to v} + \mathcal{L}_{v\to a}}{2}, \tag{6}$$

where $\tau_{\mathrm{p}} > 0$ is the temperature (set to 0.07 for the reported projector checkpoints) and $\mathcal{L}_{v\to a}$ exchanges the modalities. This symmetric InfoNCE formulation follows CLIP (Radford et al., 2021), with positives defined at the segment level. All $b' \neq b$ are negatives, including different intervals of the same video; category labels neither add positives nor filter negatives. Only the projection heads are trained. During the attack, their frozen outputs provide the similarity minimized in Eq. (3). The projector checkpoint is selected using held-out auxiliary videos and frozen before any downstream attack is generated. The auxiliary train/validation split is performed at the video level to prevent synchronized segments from the same video from appearing in both partitions. No projector parameter is updated during PGD; gradients are back-propagated through the frozen projector's input Jacobian only. The checkpoint, encoder pair, projection dimension, and temperature are kept fixed across all target models evaluated in a transfer experiment.

B.4 Perturbation Optimization

For the main transfer results, we use 50 alternating PGD steps with a uniform random start, $(\epsilon_v, \epsilon_a) = (16/255, 0.04)$, step sizes $(\eta_v, \eta_a) = (1/255, 0.002)$, intra-modal temperature $\tau = 0.07$, and $\lambda = 1$. The same perturbation budgets are used for the baseline attacks. In the ablation study, the ISC attack optimizes only the intra-modal semantics-aware contrastive loss by setting $\lambda = 0$, while the IM attack optimizes only the inter-modal misalignment loss by removing $\mathcal{L}^{\mathrm{intra}}_{s,m}$; the full AV-RIFT retains both terms. The budget ablation in the main paper additionally evaluates $(4/255, 0.01)$ and $(8/255, 0.02)$.

We initialize each perturbation uniformly within its ℓ_∞ bounds and clip it to the valid input domain. Let t index attack iterations, let $\boldsymbol{\delta}_m = (\delta_{s,m})_{s=1}^{S}$ collect the segment perturbations of modality m, and let clip_m apply the segment-wise clipping of Eq. (5) to all segments. With the video-level objective $\mathcal{L}^{\mathrm{attack}}_m = \frac{1}{S}\sum_{s=1}^{S}\mathcal{L}^{\mathrm{attack}}_{s,m}$ from Section 3.4, the alternating updates are

$$\begin{aligned} \boldsymbol{\delta}_v^{t+1} &= \mathrm{clip}_v\big(\boldsymbol{\delta}_v^t - \eta_v\,\mathrm{sign}\big[\nabla_{\boldsymbol{\delta}_v}\mathcal{L}_v^{\mathrm{attack}}(\boldsymbol{\delta}_a^t, \boldsymbol{\delta}_v^t)\big]\big)\,, \\ \boldsymbol{\delta}_a^{t+1} &= \mathrm{clip}_a\big(\boldsymbol{\delta}_a^t - \eta_a\,\mathrm{sign}\big[\nabla_{\boldsymbol{\delta}_a}\mathcal{L}_a^{\mathrm{attack}}(\boldsymbol{\delta}_a^t, \boldsymbol{\delta}_v^{t+1})\big]\big)\,. \end{aligned} \tag{7}$$

The second step uses the visual perturbation produced by the first. All clean reference features, encoders, and projection parameters remain fixed.

For input range $[l_m, u_m]$, the clipping bounds for each perturbation entry are $\max\{-\epsilon_m, l_m - x_{s,m}\}$ and $\min\{\epsilon_m, u_m - x_{s,m}\}$, with $[l_a, u_a] = [-1, 1]$ and $[l_v, u_v] = [0, 1]$. These constraints apply before video encoding. On export, visual perturbations are shared within each segment, and padded audio samples are discarded.

B.5 Attack Algorithm Summary

Algorithm 1 summarizes the complete AV-RIFT attack used to produce a transferable adversarial audio–visual input. The procedure is task agnostic: it never evaluates a downstream target model,

and it does not require target-model gradients. The only gradients used during optimization are back-propagated through the frozen surrogate encoders and the frozen projection heads. At each iteration, the objective combines the reference-based intra-modal terms and the projected inter-modal term in Eqs. (2)–(4); the perturbation update and projection onto the feasible set follow Eq. (5) and the alternating rule in Eq. (7).

Algorithm 1 Task-agnostic AV-RIFT transfer attack

Require: Clean synchronized clip X, semantic annotations $\mathcal{Y}$, frozen surrogate encoders (ψ_a, ψ_v), frozen projectors $(\mathrm{proj}_a, \mathrm{proj}_v)$, reference pools $\{\mathcal{R}^{\mathrm{sim}}_{s,m}, \mathcal{R}^{\mathrm{dis}}_{s,m}\}$, budgets (ϵ_a, ϵ_v), step sizes (η_a, η_v), and steps T

Ensure: Adversarial audio–visual clip X^{adv}

Split X into synchronized one-second segments $\{x_{s,a}, x_{s,v}\}_{s=1}^{S}$

Extract and cache clean pooled features $g_{s,m} = \mathrm{Pool}(\psi_m(x_{s,m}))$

Construct $\mathcal{R}^{\mathrm{sim}}_{s,m}$ and $\mathcal{R}^{\mathrm{dis}}_{s,m}$ from $\mathcal{Y}$

Initialize $\delta^0_{s,m} \sim \mathcal{U}[-\epsilon_m, \epsilon_m]$ for $m \in \{a, v\}$ and clip to the valid input range

for $t = 0, \ldots, T-1$ **do**

  Extract adversarial features from $x_{s,m} + \delta^t_{s,m}$ for all s, m

  Evaluate $\mathcal{L}^{\mathrm{attack}}_v$ and $\mathcal{L}^{\mathrm{attack}}_a$ using Eqs. (2)–(4)

  $\boldsymbol{\delta}^{t+1}_v \leftarrow \mathrm{clip}_v\big(\boldsymbol{\delta}^t_v - \eta_v \,\mathrm{sign}\big[\nabla_{\boldsymbol{\delta}_v} \mathcal{L}^{\mathrm{attack}}_v(\boldsymbol{\delta}^t_a, \boldsymbol{\delta}^t_v)\big]\big)$

  $\boldsymbol{\delta}^{t+1}_a \leftarrow \mathrm{clip}_a\big(\boldsymbol{\delta}^t_a - \eta_a \,\mathrm{sign}\big[\nabla_{\boldsymbol{\delta}_a} \mathcal{L}^{\mathrm{attack}}_a(\boldsymbol{\delta}^t_a, \boldsymbol{\delta}^{t+1}_v)\big]\big)$

end for

Share $\delta^T_{s,v}$ across all frames in segment s and discard audio padding

Concatenate the perturbed segments in temporal order to form X^{adv}

return X^{adv}

The clean features and reference assignments are computed once per input video and reused at every iteration, while the adversarial features are recomputed after each update. Since the loss is averaged over all segments of the video, each update jointly accounts for the temporal structure of the clip. After T steps, the segment-level perturbations are mapped back to the original waveform and frame sequence, so the exported example preserves the temporal alignment of the clean input and satisfies both modality budgets.

C Experimental Details

C.1 Evaluation Conventions

This section specifies the baseline protocol and evaluation conventions used throughout our experiments. In particular, the terms "attack success rate" (ASR) and "performance" are always computed from paired clean/adversarial predictions on the same examples; a drop in a task metric by itself is not treated as an ASR. Unless noted otherwise, attack generation uses the test split only to construct the reported examples and the target model is never queried during generation.

C.2 Baseline Attack Protocol

Shared input setting. All baseline attacks share the same numeric ranges, audio decoding, budgets, and export procedure as AV-RIFT. A clip is represented by synchronized one-second audio segments and method-specific video samples, with mono audio resampled to 16 kHz. Video tensors are in $[0, 1]$ and waveform tensors are in $[-1, 1]$. Unless a baseline is explicitly unimodal, the inactive modality is copied from the clean input. The reported budgets are $(\epsilon_v, \epsilon_a) = (16/255, 0.04)$; the same budgets are used for I2V, SoundBreak, their composition, and AV-RIFT. All surrogate encoders are frozen, and gradients are taken only with respect to the attacked input. The method-specific visual sampling and segment-to-frame mapping are stated explicitly below.

We use the notation of the main text, namely $g_{s,m} = \psi_m(x_{s,m})$ and $g^{\mathrm{adv}}_{s,m} = \psi_m(x_{s,m} + \delta_{s,m})$ for the clean and adversarial surrogate features. The representation objective shared by the two unimodal

baselines is the mean token-level cosine similarity

$$\mathcal{L}_m^{\cos} = \frac{1}{S}\sum_{s=1}^{S} \cos\big(g_{s,m}^{\mathrm{adv}}, g_{s,m}\big), \qquad m \in \{a, v\}. \tag{8}$$

The optimizer minimizes this scalar, so the clean–adversarial cosine similarity is driven down; it is not a loss that preserves the clean representation. The cosine similarity is computed for each feature token and then averaged over tokens.

I2V (visual-only). I2V (Wei et al., 2022b) studies how an image-level representation attack can transfer to video, and its central design is to disrupt visual features without requiring the downstream video model. We retain this visual-only, representation-level principle and instantiate it with the frozen ResNet-50 surrogate. Its objective is the visual specialization of Eq. (8),

$$\mathcal{L}_{\mathrm{I2V}}(\delta_v) = \mathcal{L}_v^{\cos} = \frac{1}{S}\sum_{s=1}^{S} \cos\big(g_{s,v}^{\mathrm{adv}}, g_{s,v}\big), \qquad g_{s,v}^{\mathrm{adv}} = \psi_v(x_{s,v} + \delta_{s,v}). \tag{9}$$

For the reported runs, the visual perturbation is initialized to the small constant $0.01/255$, optimized with Adam at learning rate 0.005 for 50 iterations, and projected to the visual budget and the valid pixel range after every step. The I2V sampling scheme uses eight frames sampled at rate two in each of four temporal chunks (the last deterministic temporal clip is selected); the learned segment perturbation is mapped back to the saved video with nearest temporal assignment and shared by the frames assigned to that segment. Audio remains exactly clean. The standalone I2V column in Table 1 therefore measures the transfer of a visual-only representation attack, rather than a task-loss attack.

SoundBreak (audio-only). SoundBreak (Hussain et al., 2026) systematically studies audio-only waveform attacks against trimodal models. Following its audio-only spirit, we update no visual input and use a frozen VGGish surrogate. Importantly, the SoundBreak column in this paper uses *only* the clean–adversarial cosine similarity loss

$$\mathcal{L}_{\mathrm{SoundBreak}}(\delta_a) = \mathcal{L}_a^{\cos} = \frac{1}{S}\sum_{s=1}^{S} \cos\big(g_{s,a}^{\mathrm{adv}}, g_{s,a}\big), \qquad g_{s,a}^{\mathrm{adv}} = \psi_a(x_{s,a} + \delta_{s,a}). \tag{10}$$

No cross-modal, temporal, classification, or additional distance term is added to this baseline; the cosine in Eq. (10) is minimized directly. All preprocessing, perturbation budgets, random seed, and evaluation conventions are the same as for the other reported baselines: mono 16-kHz audio, one-second synchronization, and $(\epsilon_v, \epsilon_a) = (16/255, 0.04)$. The audio-specific implementation uses a 10-second perturbation buffer initialized randomly in $[-\epsilon_a, \epsilon_a]$; a shorter clip cycles the buffer and a longer clip truncates it before addition. Adam uses learning rate 0.001, $(\beta_1, \beta_2) = (0.9, 0.999)$, and 10^{-8} numerical epsilon for the same 50-update budget. Gradients are clipped to norm 1, the perturbation is clamped to $[-0.04, 0.04]$ after every update, and a reduce-on-plateau scheduler (factor 0.1, patience 5) is used. The perturbed waveform is finally normalized and clipped to $[-1, 1]$ before it is combined with the video stream; the visual stream remains clean. Thus, differences between SoundBreak and I2V in Table 1 arise from the attacked modality and its waveform parameterization, not from importing a different loss family.

I2V + SoundBreak. The combined baseline is our composition of the two attacks above, not a separate method proposed in either paper. The visual phase is completed first, followed by the audio phase; each phase compares its adversarial feature with the clean feature and uses no target-model output. The two deltas are then applied to the same temporally aligned clip and exported together. The combination therefore does not perform a joint alternating optimization and does not re-optimize either modality after the other delta is inserted, and no cross-modal loss is used. The combined column should therefore be interpreted as an independent-composition baseline rather than as a variant of AV-RIFT's joint objective.

TMA. TMA (Zhang et al., 2025) introduces a temporal-and-modality view of audio–visual vulnerability: the temporal term disrupts the consistency of features across time, while the modality term reduces the audio–visual agreement. In schematic form, its minimized objective can be written as

$$\mathcal{L}_{\mathrm{TMA}} = \mathcal{L}_{\mathrm{primary}} + \lambda_{\mathrm{TIA}}\mathcal{L}_{\mathrm{TIA}} + \lambda_{\mathrm{MMA}}\mathcal{L}_{\mathrm{MMA}}, \tag{11}$$

where $\mathcal{L}_{\text{primary}}$ is the source-task classification loss in the original task-aware formulation. For the temporal term, our implementation computes the variance of per-segment feature means, e.g., $\mathcal{L}_{\text{TIA}}^{m} = \text{Var}_s[\text{mean}(g_{s,m}^{\text{adv}})]$; the modality term is a cosine agreement loss between the aligned ResNet/VGGish audio and visual features. TMA is included only on AVER, where its task-specific formulation is defined. TMA can be instantiated in two modes. The task-aware mode uses the classification loss of a source AVER model as $\mathcal{L}_{\text{primary}}$, which requires white-box access to that model and is therefore not a black-box transfer attack. The proxy mode replaces $\mathcal{L}_{\text{primary}}$ with the clean–adversarial feature cosine similarity while retaining the temporal and modality terms. The reported TMA results in Table 4 use the task-aware mode with R + A + Concat as the source model. In both modes, we optimize TMA with MI-FGSM using 50 steps, momentum decay 1, step sizes $(\eta_v, \eta_a) = (1/255, 0.002)$, and the same $(16/255, 0.04)$ budgets; the optional input transformation is sampled once per step. Except for the TMA-setting comparison in Table 4, where AV-RIFT additionally uses the classification loss of the same source model, AV-RIFT and the I2V + SoundBreak baseline never use target-model logits, predicted labels, or gradients during attack generation.

Reproducibility controls. For every method we fix the random seed, segment order, clean checkpoint, and sample identifiers across the clean and adversarial passes. Perturbations are clipped before media export, and the exported audio/video is evaluated with the same preprocessing as the clean input. Consequently, differences in the next subsection reflect the attack and not a changed query set, gallery, or input geometry.

C.3 Task-level Evaluation and Attack Success Criteria

Paired evaluation and conditioning. Let $\mathcal{I}$ be the set of examples that have both a clean and an adversarial prediction from the same target checkpoint. We preserve the sample/question/frame identifiers and prediction order when pairing the two passes. Task performance is computed on all valid records in $\mathcal{I}$, whereas ASR is conditional on a task-specific clean-success set. Thus, an example that the target model already gets wrong on clean input is never counted as a successful untargeted attack. For a classification output unit i,

$$\text{Acc} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbf{1}[\hat{y}_i^{\text{clean}} = y_i], \qquad \text{ASR}_{\text{cls}} = \frac{\sum_{i \in \mathcal{I}} \mathbf{1}[\hat{y}_i^{\text{clean}} = y_i] \mathbf{1}[\hat{y}_i^{\text{adv}} \neq y_i]}{\sum_{i \in \mathcal{I}} \mathbf{1}[\hat{y}_i^{\text{clean}} = y_i]}. \tag{12}$$

AVEL and AVER. For AVEL, an output unit is a video segment and y_i is its event label; the reported accuracy and ASR are therefore segment-level quantities. For AVER, an output unit is a video clip and the same formula is applied to its event class. In both cases, "success" means that a clean-correct label changes to an incorrect label; a change to a particular target class is not required.

AVS. For AVS we compute the native mean IoU over the valid output frames. The primary conditional ASR is the exact threshold event

$$\text{ASR}_{\text{AVS}} = \frac{\sum_i \mathbf{1}[\text{IoU}_i^{\text{clean}} \geq 0.5] \mathbf{1}[\text{IoU}_i^{\text{adv}} < 0.1]}{\sum_i \mathbf{1}[\text{IoU}_i^{\text{clean}} \geq 0.5]}. \tag{13}$$

The inequalities are intentional: clean IoU equal to 0.5 qualifies for the denominator, while adversarial IoU equal to 0.1 is not a success. Empty foreground frames follow the evaluator's background-IoU convention. The unconditional fraction of adversarial frames below 0.1 is retained only as a diagnostic and is not substituted for the primary clean-conditioned ASR.

AVR. For AVR, the query set and gallery are fixed to the same VEGAS evaluation fold in both passes. Every gallery item with the same event category as the query is relevant (category-level, not exact-instance, relevance). We rank the clean gallery separately for audio-to-visual (A2V) and visual-to-audio (V2A), compute AP for each query, and report

$$\text{mAP} = \tfrac{1}{2} \left(\text{mAP}_{\text{A2V}} + \text{mAP}_{\text{V2A}} \right). \tag{14}$$

For attack evaluation, adversarial queries are compared with the unchanged clean gallery. The directional ASR is the fraction of queries whose clean top-ranked item is relevant but whose adversarial top-ranked item is irrelevant:

$$\text{ASR}_d = \frac{\#\{i : r_{i,d}^{\text{clean}} = 1, r_{i,d}^{\text{adv}} = 0\}}{\#\{i : r_{i,d}^{\text{clean}} = 1\}}, \quad d \in \{\text{A2V}, \text{V2A}\}. \tag{15}$$

The reported AVR ASR is the arithmetic mean of the two directional ASRs, not a ratio formed by pooling their denominators. This prevents one retrieval direction from dominating merely because it has more clean top-1 successes.

AVQA. For AVQA, the output unit is a question and accuracy uses exact answer-label match (equivalently, exact match after the model's answer normalization). If question-level clean and adversarial records are available, its ASR follows Eq. (12) with "label" replaced by the answer. A question that is clean-incorrect is excluded from the ASR denominator, even if its adversarial answer changes. ASR is computed only from paired question-level predictions and is never inferred from an aggregate accuracy drop.

Aggregation across models. For each task, we first compute the metric and the conditional ASR separately for every target model, then take the unweighted arithmetic mean across models within that task. We do not pool segment, frame, clip, question, or retrieval records across different tasks, and we do not form a cross-task ASR average from incompatible denominators. Clean and adversarial values in Table 1 therefore share model checkpoints and evaluation records, while the task-level averages remain comparable without allowing a task with more output units to dominate the summary.