Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Candidate-Conditioned Ontology-Guided Audio-Visual Scene Classification under Ambiguous Acoustic Evidence

Hoang Giang Vu

hoang.vu@univ-tours.fr

University of Tours

Research Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11578224/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Candidate-Conditioned Ontology-Guided Audio-Visual Scene Classification under Ambiguous Acoustic Evidence

Hoang Giang Vu[1*]

[1*]LIFAT, University of Tours, Tours, France.

Corresponding author(s). E-mail(s): hoang.vu@univ-tours.fr;

Abstract

Audio-visual scene classification can benefit from complementary modalities, but urban acoustic evidence is often ambiguous and less reliable than visual information for direct scene recognition. We investigate whether such weak acoustic evidence can improve a strong visual classifier when interpreted through explicit semantic knowledge. We propose a training-free, candidate-conditioned ontology-guided framework in which CLIP provides visual predictions and a Top-3 candidate set, while AST extracts temporally persistent acoustic events. A domain-informed ontology converts these events into semantic evidence for the competing scene hypotheses, providing an interpretable event-to-concept-to-candidate reasoning path. Candidate-conditioned rules further refine ambiguous cases while preserving reliable visual predictions. On the 6000-clip development set, the proposed framework improves CLIP accuracy from 78.93% to 81.12% (+2.19 pp), while on the independent 3500-recording evaluation set it improves accuracy from 77.03% to 78.97% (+1.94 pp). The ontology alone also provides consistent gains, reaching 79.73% and 77.80%, respectively. These results show that structured acoustic knowledge can improve visual scene classification while providing an explicit explanation of when and why acoustic evidence modifies a prediction. Git: https://github.com/jangvu/stvd-avsc

Keywords: Audio-visual scene classification, Multimodal fusion, Ontology-guided reasoning,Zero-shot classification

1 Introduction

Multimodal learning exploits complementary information from heterogeneous sources such as audio and vision [1]. In urban scene understanding, vision captures objects and spatial context, whereas audio provides cues from vehicles, speech, crowds, and environmental events [2]. The TAU Audio-Visual Urban Scenes dataset supports this task across ten urban categories [3], and previous work has shown that combining both modalities can improve scene recognition [4].

However, the two modalities are not equally reliable for every sample. Urban sounds are often weak, overlapping, and shared across multiple environments; events such as engines, speech, and doors may therefore provide ambiguous scene evidence [5, 6]. Consequently, global fusion can allow unreliable acoustic information to disturb otherwise confident visual predictions [7, 8]. Recent pretrained models provide strong representations for both modalities: CLIP enables zero-shot visual recognition [9], CLAP supports audio–text classification [10], and AST together with the AudioSet taxonomy represents audio through semantic sound events [11, 12].

Existing audio-visual scene classification methods largely treat audio and vision as parallel

prediction streams and combine them through global feature- or decision-level fusion. This is less suitable when one modality is substantially weaker than the other, and acoustic events are rarely used as structured semantic evidence conditioned on a compact visual candidate set. On the development subset, CLIP achieves 78.93% top-1 accuracy while the correct class appears within its top-three predictions for 99.35% of recordings, indicating that many remaining errors arise from candidate selection rather than candidate generation. We therefore use audio as selective semantic evidence instead of as an equally reliable scene classifier.

We propose a training-free, candidate-conditioned ontology-guided framework in which CLIP generates visual top-K candidates, AST extracts persistent acoustic events, and a domain-informed ontology maps these events to semantic scene cues. Candidate-conditioned rules then determine whether the acoustic evidence is sufficiently relevant to retain or correct the visual prediction. The resulting decision process is interpretable because each intervention can be traced from detected events through semantic concepts to competing scene hypotheses.

The main contributions are:

- a training-free audio-visual framework that uses a weaker acoustic modality selectively to improve a stronger visual classifier;
- a domain-informed acoustic ontology that maps sound events to semantic scene cues;
- a candidate-conditioned reasoning mechanism that refines visual top-K predictions while protecting reliable visual decisions; and
- an interpretable decision process in which corrections are traceable to explicit acoustic events, ontology relations, and candidate scenes.

2 Related Work

2.1 Audio-Visual Scene Classification

The TAU Audio-Visual Urban Scenes dataset provides synchronized urban audio and video across ten scene categories [3]. The DCASE 2021 challenge subsequently established a common AVSC benchmark, with its strongest submissions combining pretrained representations and multimodal fusion [13]. Naranjo-Alcazar et al. investigated early and late audio-visual fusion [4], while Okazaki et al. incorporated CLIP representations into confidence-level multimodal fusion [14]. More recent methods employ hierarchical semantic attention [15] and multi-scale temporal cross-modal interaction [16]. These approaches predominantly learn or globally combine multimodal representations. In contrast, the present work treats vision as the primary decision source and allows acoustic evidence to intervene only when it is semantically relevant to the current candidate set [8].

2.2 Scene–Event and Knowledge-Guided Reasoning

Scene recognition can benefit from the semantic relationships between acoustic scenes and their constituent events. Tonami et al. jointly modeled sound events and acoustic scenes through multitask learning [17], while Hou et al. introduced explicit scene–event relations through a prior relation matrix [18], event-relational graphs [19], and cooperative scene–event modeling [6]. Semantic reasoning has also been applied directly to AVSC through fine-grained event–object alignment and semantic fusion [20]. Existing methods primarily learn scene-event relationships or integrate them globally into multimodal representations. In contrast, the proposed framework uses an explicit acoustic ontology to interpret frozen sound-event predictions only within the compact hypothesis set generated by the visual classifier. Thus, acoustic evidence acts as candidate-specific semantic support rather than as an independent scene predictor or uniformly fused modality.

3 Proposed Method

3.1 Design Motivation and Model Selection

3.1.1 Zero-Shot Scene Recognition

The proposed framework is motivated by the observation that scene-defining acoustic information is often weak, overlapping, and non-unique in the TAU Audio-Visual Urban Scenes recordings [3]. The recordings contain complex urban acoustic mixtures in which several sound sources may occur simultaneously, relevant events can be

distant or intermittent, and the same sound can occur in several different scene classes. For example, engines, doors, speech, music, and general vehicle sounds can occur in bus, metro, tram, metro-station, and street environments, while outdoor human activity may be shared by park, pedestrian-street, and public-square recordings. The audio recordings therefore provide useful semantic cues but do not always contain a clean scene-defining sound, making direct scene classification substantially more difficult than visual scene recognition.

To quantify this difference, three pretrained models are first considered for direct scene-level recognition: CLIP [9], LAION-CLAP [10], and Qwen2.5-Omni-3B. CLIP performs zero-shot visual classification by comparing sampled video frames with textual descriptions of the ten TAU scene classes. LAION-CLAP performs zero-shot audio classification by comparing each recording with textual scene prompts in a joint audio–text embedding space. Qwen2.5-Omni-3B is evaluated as a general-purpose audio-language model by providing the audio recording and a constrained instruction requiring one of the ten scene labels. On the common 6000-clip subset, CLIP achieves 78.93% top-1 accuracy, while LAION-CLAP and Qwen2.5-Omni-3B achieve 48.27% and 38.10%, respectively (Tab 1). The substantial gap between visual and audio-only recognition indicates that directly predicting the complete scene from the acoustic modality is unreliable under the complex recording conditions of this dataset. However, the audio-only results remain well above random prediction, suggesting that the recordings still contain useful partial acoustic information.

This observation motivates the proposed strategy: the visual modality generates a small set of plausible scene hypotheses, while the acoustic modality is used to provide supporting or contradictory evidence among those candidates instead of independently deciding the final scene.

3.1.2 Visual Candidate Generation

The visual branch uses CLIP ViT-B/32 [9] as the primary scene classifier and candidate generator.

Eight frames are sampled from approximately equal temporal regions of each video, and the zero-shot probabilities obtained from the ten fixed scene prompts are averaged across the sampled frames.

Table 1 Comparison of model choices for scene recognition and event-assisted fusion.

Method	Mod.	Acc. (%)	Macro-F1
Qwen2.5-Omni-3B	A	38.10	0.2992
LAION-CLAP	A	48.27	0.4371
CLIP	V	78.93	0.7877
CLIP+PANNs+Event Map	A+V	79.02	0.7887
CLIP+AST+Event Map	A+V	**79.42**	**0.7925**

A: audio; V: visual. Best fusion weights are 0.85/0.15 for CLIP/PANNs and 0.80/0.20 for CLIP/AST.

For video v, let $p_v(c)$ denote the averaged CLIP probability for scene class c.

The visual candidate set is defined as

$$\mathcal{C}_K(v) = \mathrm{TopK}_c \; p_v(c), \tag{1}$$

where $K = 3$ in the proposed framework.

Although CLIP achieves 78.93% top-1 accuracy, the ground-truth scene occurs within the top-three candidates for 99.35% of the evaluated recordings.

This difference shows that most CLIP errors occur because the correct candidate is not ranked first rather than because the correct class is completely absent from the visual hypothesis set.

The acoustic branch is therefore designed to assist primarily with candidate selection inside $\mathcal{C}_3(v)$.

3.1.3 Acoustic Event Extraction and Distribution

The acoustic branch represents each recording using intermediate sound events rather than directly predicting the scene label. Both PANNs [21] and AST [11] were investigated because they provide event-level predictions related to the AudioSet ontology [12]. To select the acoustic event extractor, PANNs and AST were first evaluated using the same simple event-to-scene mapping combined with CLIP. As shown in Table 1, CLIP+PANNs achieves 79.02% accuracy and 0.7887 macro-F1, whereas CLIP+AST reaches 79.42% accuracy and 0.7925 macro-F1. AST is therefore retained as the acoustic event extractor for the subsequent ontology-guided framework.

Table 2 Representative high-occurrence AST events across TAU scenes. Values in parentheses indicate occurrence percentage.

Scene	Representative AST events (%)
Airport	Speech (96.50); Music (19.67); Vehicle (15.00); Sliding door (13.83)
Bus	Vehicle (91.67); Speech (74.17); Car (44.67); Music (23.83)
Metro	Speech (88.83); Vehicle (82.33); Music (21.17); Sliding door (18.83)
Metro station	Speech (79.67); Vehicle (56.17); Music (35.83); Sliding door (20.50)
Park	Speech (49.00); Bird (18.83); Vehicle (18.33); Music (10.17)
Public square	Speech (90.00); Vehicle (37.00); Music (14.17); Car (5.33)
Shopping mall	Speech (98.00); Music (47.67); Sliding door (14.50); Vehicle (12.33)
Pedestrian street	Speech (96.67); Music (24.83); Vehicle (14.67); Sliding door (6.83)
Street traffic	Vehicle (93.67); Car (63.67); Speech (31.17); Music (4.50)
Tram	Speech (84.00); Vehicle (71.17); Music (23.17); Sliding door (22.17)

After selecting AST, it is applied across the 6000-recording development subset to examine the acoustic characteristics of each scene category. The detected AudioSet events are aggregated by scene, and Table 2 presents representative high-occurrence events from the selected acoustic-event set.

The updated AST event distributions reveal substantial acoustic overlap across the TAU scene classes, showing that individual events are often insufficient for direct scene classification. For example, *Vehicle* appears in 93.67% of street-traffic and 91.67% of bus recordings, but is also frequent in metro (82.33%), tram (71.17%), and metro-station (56.17%) scenes. Likewise, *Car* occurs in street traffic (63.67%) and bus (44.67%), while also appearing in tram (20.50%) and metro (11.50%) recordings. The *Bird* event is most prominent in park recordings (18.83%), but is also detected in several other classes, and even the more specific *Bus* event is not exclusive to bus scenes. These observations indicate that acoustic events provide useful but ambiguous scene evidence, motivating the use of multiple events together with their confidence and temporal persistence rather than relying on a single highest-scoring event.

Each audio recording is converted to 16-kHz mono and divided into 3-s segments, with the eight highest-scoring AST events retained for each segment. For event e, temporal persistence is defined as

$$\rho_e = \frac{n_e}{N}, \tag{2}$$

where n_e is the number of segments containing event e and N is the total number of segments.

The resulting representation stores the event confidence, occurrence frequency, and persistence across the recording. These corpus-level event patterns provide the empirical basis for constructing the acoustic ontology introduced in Sec. 3.2, where overlapping event detections are interpreted through broader semantic scene relationships rather than direct event-to-scene assignments.

3.2 Semantic Acoustic Reasoning

3.2.1 Corpus-Informed Acoustic Ontology

The corpus-level AST analysis shows that directly assigning a detected acoustic event to a single TAU scene is unreliable because many sounds occur across several environments. We therefore construct an acoustic-scene ontology in which selected AST events are grouped into broader acoustic concepts and connected to semantically related TAU scene classes. The ontology is informed by the observed TAU event distributions and uses 22 AudioSet-related event mappings covering road transport, rail transport, transport access, natural, outdoor, and air-transport context. Table 3 summarizes the acoustic vocabulary and its semantic use.

Conceptually, the ontology follows the semantic path

$$\begin{aligned}\text{AST event} &\rightarrow \text{acoustic concept} \\ &\rightarrow \text{semantically related candidate scene.}\end{aligned} \tag{3}$$

rather than imposing a direct one-event–one-scene assignment.

For acoustic event e and candidate scene c, its semantic contribution is

$$E(e, c) = s_e \, \rho_e \, w(e, c), \tag{4}$$

Table 3 Acoustic concepts and AST events used in the ontology.

Concept	AST / AudioSet events	Related scenes
Road transport	Vehicle; Car; Motor vehicle; Engine; Vehicle horn; Air brake; Bus	Bus; Metro; Metro station; Street traffic; Tram
Transport access	Sliding door; Door	Bus; Metro; Metro station; Tram
Rail transport	Train; Rail transport; Subway/metro; Wheel squeal; Railroad car	Metro; Metro station; Tram
Air transport	Aircraft; Aircraft engine; Jet engine; Fixed-wing aircraft	Airport
Natural context	Bird; Bird vocalization/bird call/bird song	Park; Public square; Pedestrian street
Outdoor context	Outside urban	Public square; Pedestrian street; Street traffic; Park

where s_e denotes the aggregated AST confidence, ρ_e denotes temporal persistence, and $w(e, c)$ denotes the semantic relation strength. The total ontology evidence for candidate c is

$$O(c) = \sum_{e \in \mathcal{E}} E(e, c), \tag{5}$$

where $\mathcal{E}$ denotes the selected ontology event vocabulary.

The relation weight $w(e, c)$ allows an event to contribute differently to multiple related scenes rather than imposing a binary association. For example, *Car*, *Motor vehicle (road)*, and *Engine* provide shared road-transport evidence, while more specific events such as *Bus* provide stronger support for their corresponding scene. The same principle is applied to rail, access, natural, outdoor, and air-transport events. Only classes contained in the CLIP top-three candidate set receive ontology evidence, while the fusion and protection rules are described in Sec. 3.3.

3.2.2 Candidate-Conditioned Reasoning for Similar Environments

The basic ontology captures global event-to-scene relations, but several TAU classes remain difficult because closely related environments share similar acoustic evidence. Candidate-conditioned reasoning therefore interprets events relative to the current CLIP top-three configuration.

For $\{\text{bus}, \text{metro}, \text{tram}\}$, road-related cues such as motor-vehicle, engine, truck, acceleration, and passing-vehicle events support *bus*, whereas train and railroad-car events provide rail-family evidence. Generic rail evidence is not used to choose directly between *metro* and *tram*; the visual scores retain that decision.

For $\{\text{park}, \text{pedestrian street}, \text{street traffic}\}$, park support requires persistent bird activity together with a quiet context and the absence of strong traffic evidence. Silence alone is not sufficient.

For $\{\text{metro}, \text{metro station}, \text{tram}\}$, persistent human context from speech or music must be accompanied by independent rail or access evidence before correcting to *metro station*. Speech, music, or door sounds alone therefore cannot trigger a station correction.

These rules allow the same acoustic event to have different importance depending on the competing visual candidates, rather than applying one universal sound-to-scene mapping.

3.3 Multimodal Fusion and Selective Intervention

The ontology converts the retained AST events into semantic evidence for each scene contained in the CLIP top-three candidate set.

After normalization, the visual and ontology evidence are combined as

$$F(c) = p_v(c) + \lambda \widetilde{O}(c), \tag{6}$$

where $p_v(c)$ is the CLIP probability, $\widetilde{O}(c)$ is the normalized ontology evidence, and λ controls the contribution of the acoustic modality.

A higher fused score alone is not sufficient to replace the CLIP prediction because weak or ambiguous acoustic events may otherwise introduce harmful corrections. The proposed method

therefore applies a protection gate in which an alternative candidate must provide sufficient ontology evidence, a minimum evidence advantage over the current prediction, and a sufficiently small visual confidence gap.

For the basic ontology, AST events with confidence below 0.08 are discarded. The ontology contribution is weighted by $\lambda = 0.4$, with a minimum ontology evidence of 0.15, a minimum evidence advantage of 0.05, and a CLIP protection margin of 0.2. Ontology evidence of at least 0.60 is considered strong and may permit a correction beyond the standard visual protection margin.

The candidate-conditioned rules use separate thresholds because acoustic cues become more informative once the competing visual hypotheses are known. For the unordered set $\{\text{bus}, \text{metro}, \text{tram}\}$, road or rail cues must satisfy a confidence threshold of 0.015 and persistence of 0.25. At least two road cues are required to support *bus*, while one or more rail cues can challenge *bus*; the CLIP score-gap limit is 0.20 and increases to 0.45 under stronger corroborating evidence. For $\{\text{metro}, \text{metro station}, \text{tram}\}$, speech or music must be supported by rail or access evidence before correcting to *metro station*. Thus, ambiguous acoustic events affect the prediction only within an appropriate candidate configuration and with corroborating evidence.

The fusion stages are applied sequentially: the basic ontology is evaluated first, and the candidate-conditioned stage is applied only when the ontology preserves the original CLIP prediction. This prevents reliable ontology corrections from being overwritten. When the acoustic evidence is weak, generic, contradictory, or insufficient to satisfy the relevant conditions, no additional correction is made and the current prediction is retained.

4 Experiments and Discussion

4.1 Dataset and Evaluation Protocol

Experiments are conducted on the TAU Audio-Visual Urban Scenes 2021 dataset [3]. A class-balanced subset of 6000 synchronized audio–video recordings, with 600 samples from each of the ten scene categories, is used for method development. This development subset is used to analyze CLIP errors, construct the acoustic ontology, select AST event groups, and determine the candidate-conditioned rules and associated thresholds.

After the ontology structure, semantic relations, thresholds, and candidate rules are fixed, the complete framework is evaluated on an independent set of 3500 previously unseen audio–video recordings. These evaluation samples are not used during ontology construction or rule development. The same evaluation set is used for CLIP, the basic ontology configuration, and the final candidate-conditioned system, enabling direct paired comparison.

The proposed pipeline (Fig. 1) uses CLIP ViT-B/32 [9] as the primary visual classifier and candidate generator. AST [11] is used to extract event-level acoustic evidence, which is subsequently mapped to semantic scene concepts through the proposed ontology. Two multimodal configurations are evaluated: *CLIP+AST+Ontology*, which applies global ontology-guided acoustic reasoning, and the full proposed method, which further introduces candidate-conditioned rules for resolving specific visual ambiguities. Top-1 accuracy and macro-F1 are reported as the primary classification metrics. Prediction-level changes are additionally analyzed using wrong-to-correct (W→C), correct-to-wrong (C→W), and wrong-to-wrong (W→W) transitions.

Table 4 Classification results on the development and independent evaluation sets.

Dataset	Method	Acc. (%)	Macro-F1
Development	CLIP	78.93	0.7877
	CLIP+AST+Ontology	79.73	0.7959
	Proposed	**81.12**	**0.8084**
Evaluation	CLIP	77.03	0.7675
	CLIP+AST+Ontology	77.80	0.7751
	Proposed	**78.97**	**0.7858**

4.2 Evaluation Results

Table 4 summarizes results on the 6000-AV development set and the independent 3500-AV evaluation set. All ontology relations, event groups, fusion weights, thresholds, candidate-conditioned rules, and protection conditions were determined exclusively on the development set and fixed before evaluation, with no evaluation samples used for model selection or rule modification.

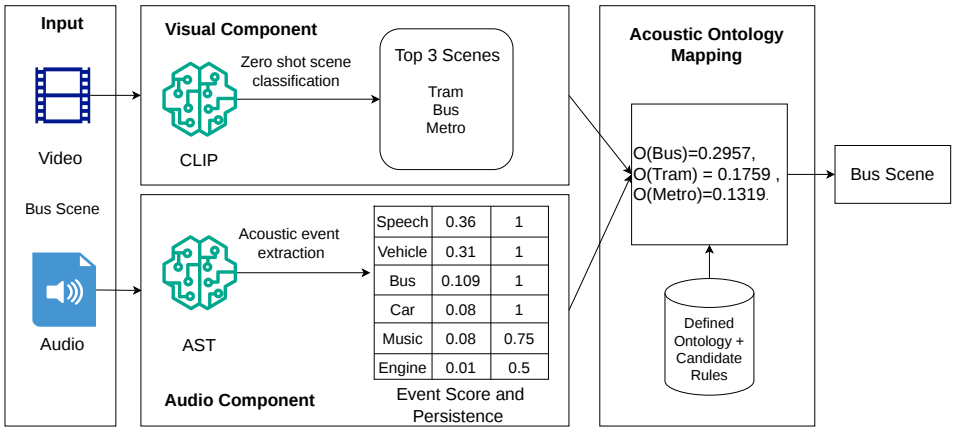


Fig. 1 Pipeline of the proposed candidate-conditioned ontology-guided audio-visual scene classification framework.

On the development subset, ontology-guided reasoning improves the CLIP baseline from 78.93% to 79.73% accuracy, while the complete candidate-conditioned framework reaches 81.12% accuracy and 0.8084 macro-F1, corresponding to a 2.19 percentage-point gain over CLIP. The same trend is preserved on the independent evaluation set, where CLIP achieves 77.03% accuracy and 0.7675 macro-F1, while the proposed method reaches 78.97% accuracy and 0.7858 macro-F1, corresponding to a 1.94 percentage-point improvement over CLIP. These results indicate that the benefit of candidate-conditioned acoustic reasoning is not restricted to the development data.

For historical context, the official DCASE 2021 Task 1B audio-visual baseline reports 77.0% accuracy under the official DCASE development protocol. Across the combined 9500 recordings used in our development and independent evaluation sets, the proposed method achieves an approximate weighted accuracy of 80.33%.

Next, to examine how acoustic reasoning modifies the predictions, Table 5 reports the transitions introduced at each reasoning stage. For both configurations, prediction transitions are measured relative to the original CLIP predictions. W→C denotes an incorrect prediction corrected by the corresponding stage, C→W denotes a correct prediction changed to an incorrect one, and W→W denotes a change from one incorrect class to another.

The basic ontology exhibits a conservative intervention pattern, producing a net gain of 48 correct predictions on the development set and 27 on the independent evaluation set. After candidate-conditioned reasoning is included, the complete framework becomes more active: relative to the CLIP baseline, it produces 169 W→C, 38 C→W, and 26 W→W changes on the development set, yielding a net gain of 131 recordings. On the evaluation set, it produces 94 W→C, 26 C→W, and 25 W→W changes, corresponding to a net gain of 68 recordings. These results show that conditioning acoustic evidence on the competing visual candidates enables substantially more useful corrections while retaining protection against excessive acoustic intervention. For applications prioritizing reliability, a more conservative setting with minimum event confidence 0.10, $\lambda = 0.40$, and CLIP protection margin 0.20 reduces harmful interventions to only 1 C→W case on development, while retaining 47 W→C and 3 W→W changes. This configuration still achieves 79.70% accuracy and 0.7957 macro-F1, showing that the framework can trade a small amount of correction activity for a safer intervention profile.

Table 5 Prediction changes introduced by ontology and candidate-conditioned reasoning.

Method	W→C	C→W	W→W
Dev.: Ontology	51	3	4
Dev.: Candidate rules	169	38	26
Eval.: Ontology	33	6	4
Eval.: Candidate rules	94	26	25

Table 6 Ablation results on the development (D) and evaluation (E) sets.

Configuration	D (%)	E (%)
Full ontology	79.73	77.80
w/o persistence	79.85	77.80
w/o CLIP protection	79.70	77.74
Full candidate	**81.12**	**78.97**
w/o transport rules	80.17	78.14
w/o park rule	81.10	78.97
w/o station rules	80.70	78.63

Additional ablation experiments evaluate temporal persistence, the ontology-level CLIP protection gate, and the candidate-conditioned rule families. As shown in Table 6, removing persistence increases development accuracy from 79.73% to 79.85%, while evaluation accuracy remains unchanged at 77.80%. This suggests that persistence is less useful for the present short recordings, but may become more important for longer or acoustically complex scenes with recurring events. In contrast, removing the CLIP protection gate reduces accuracy from 79.73% to 79.70% on development and from 77.80% to 77.74% on evaluation, showing that the gate helps prevent weak acoustic evidence from overriding confident visual predictions.

The candidate-conditioned ablation shows a clearer dependence on the semantic rule families. Removing the transport rules reduces accuracy from 81.12% to 80.17% on development and from 78.97% to 78.14% on evaluation, representing the largest degradation among the tested rule families. Removing the metro-station rules also decreases performance, to 80.70% and 78.63%, respectively. In contrast, removing the park rule has only a negligible effect, yielding 81.10% on development and no measurable accuracy change on evaluation. Finally, disabling all candidate-conditioned rules reduces the method to the basic ontology, with accuracies of 79.73% and 77.80%. Overall, these results indicate that the main benefit of candidate-conditioned reasoning arises from resolving visually ambiguous transport and station-related environments, while the park-specific rule contributes only marginally on the present datasets. Additional configurations and implementation details are provided in the public repository.

We also examine the class-wise effects of acoustic reasoning. Table 7 shows that its contribution

Table 7 Per-class accuracy gains (pp) over the CLIP baseline. O denotes ontology-guided reasoning, while OC denotes the full ontology with candidate-conditioned reasoning. Bold indicates the largest per-class gain in each column.

	Development		Evaluation	
Class	O	OC	O	OC
Airport	0.00	0.00	0.00	0.00
Bus	+5.00	+17.17	+6.86	+18.29
Metro	+0.17	+0.33	+0.29	0.00
Metro station	0.00	+4.17	+0.29	+3.71
Park	+0.33	+0.50	+0.29	+0.29
Public square	0.00	0.00	0.00	-0.29
Shopping mall	0.00	0.00	0.00	0.00
Pedestrian street	0.00	0.00	0.00	0.00
Street traffic	+2.33	+2.33	+1.71	+1.71
Tram	+0.17	-2.67	-1.43	-4.29

is strongly class dependent. The ontology-only configuration (O) improves overall accuracy by 0.80 pp on development and 0.77 pp on evaluation, while the full ontology with candidate-conditioned reasoning (OC) increases these gains to 2.19 pp and 1.94 pp, respectively. The largest improvements are observed for *bus*, where OC improves the CLIP baseline by 17.17 pp on development and 18.29 pp on evaluation. *Metro station* also benefits clearly, with gains of 4.17 pp and 3.71 pp, while *street traffic* improves by 2.33 pp and 1.71 pp. These results indicate that candidate-conditioned acoustic reasoning is particularly effective for visually ambiguous transport and station-related environments. In contrast, *tram* remains challenging, with OC reducing accuracy relative to CLIP by 2.67 pp on development and 4.29 pp on evaluation.

For classes such as *airport*, *public square*, *shopping mall*, and *pedestrian street*, acoustic reasoning provides little or no additional gain. This does not imply an absence of detected acoustic events; rather, the available evidence is often generic, overlapping, or insufficiently discriminative to justify a reliable correction. Indoor scenes may share events such as speech, music, door activity, and background noise, while outdoor classes such as *public square* and *pedestrian street* contain similar human-activity and urban sounds. In these cases, the protection mechanism retains the original CLIP prediction when the acoustic evidence is weak or ambiguous, which is consistent with the selective-intervention objective of the framework. The main negative case is *tram*, where road-transport cues such as vehicle, car,

and engine sounds can overlap with *bus*, causing candidate-conditioned reasoning to favor *bus* when sufficiently distinctive rail-specific evidence is absent.

Overall, the results support the central hypothesis that a comparatively weak acoustic modality can still improve a stronger visual classifier when used as selective semantic evidence rather than as an independent scene predictor. The ontology provides conservative and reliable acoustic filtering, while candidate-conditioned reasoning introduces a stronger correction mechanism that yields the larger accuracy gain. Together, these components improve recognition performance while retaining an interpretable reasoning path from detected acoustic events to semantic concepts and competing visual scene hypotheses.

5 Limitations

The proposed framework has three main limitations. First, acoustic evidence is not equally discriminative across scene categories, particularly when foreground events are weak, overlapping, or shared across acoustically similar environments. Second, the ontology and candidate-conditioned rules are domain dependent, since event–scene relations, thresholds, and candidate configurations may require adaptation to new scene taxonomies or acoustic domains. Third, the method depends on the quality of the CLIP candidate set and cannot recover a class that is absent from the visual top-K hypotheses. Although the fixed framework generalizes to an independent evaluation set, its ontology structure and decision thresholds are manually defined from development data. Future work will therefore explore adaptive candidate selection and automatic learning of event–scene relations while preserving the framework's explicit and interpretable reasoning.

6 Conclusion

This work shows that a comparatively weak acoustic modality can improve a stronger visual scene classifier when sound events are interpreted as structured semantic evidence rather than used for direct scene prediction. The proposed framework uses CLIP to generate visual candidates, AST to extract persistent sound events, and a defined acoustic ontology to organize overlapping events into meaningful scene-related concepts. The results indicate that the quality and semantic relevance of the detected sound events are important: distinctive events provide useful complementary evidence, whereas generic or overlapping sounds offer limited discrimination. The ontology therefore acts as a semantic layer, while the candidate-conditioned rules further exploit acoustic evidence within specific visual candidate configurations and prevent weak cues from unnecessarily altering reliable visual predictions. This combination improves accuracy from 78.93% to 81.12% on the development set and from 77.03% to 78.97% on the independent evaluation set, with macro-F1 increasing from 0.7675 to 0.7858 on the evaluation set. Overall, the findings demonstrate that explicitly defined event-to-concept relations together with candidate-specific acoustic reasoning can make weak audio evidence useful, interpretable, and complementary to strong visual recognition without task-specific neural-network retraining.

Declarations

Funding. The author receives financial support through a funded doctoral position. However, the research presented in this manuscript was conducted independently of the funded doctoral project and received no dedicated research funding.

References

[1] Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 423–443 (2019)

[2] Zhu, H., Luo, M.-D., Wang, R., Zheng, A.-H., He, R.: Deep audio-visual learning: A survey. Mach. Intell. Res. **18**, 351–376 (2021)

[3] Wang, S., Mesaros, A., Heittola, T., Virtanen, T.: A curated dataset of urban scenes for audio-visual scene analysis. In: Proc. IEEE ICASSP, pp. 626–630 (2021)

[4] Naranjo-Alcazar, J., Perez-Castanos, S., Lopez-Garcia, A., Zuccarello, P., Cobos, M., Ferri, F.J.: Squeeze-excitation convolutional recurrent neural networks for audio-visual

scene classification. In: Proc. DCASE Workshop, pp. 16–20 (2021)
[5] Ding, B., Zhang, T., Wang, C., Liu, G., Liang, J., Hu, R., Wu, Y., Guo, D.: Acoustic scene classification: A comprehensive survey. Expert Syst. Appl. **238**, 121902 (2024)
[6] Hou, Y., Kang, B., Mitchell, A., Wang, W., Kang, J., Botteldooren, D.: Cooperative scene-event modelling for acoustic scene classification. IEEE/ACM Trans. Audio Speech Lang. Process. **32**, 68–82 (2024)
[7] Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multimodal transformers robust to missing modality? In: Proc. CVPR, pp. 18177–18186 (2022)
[8] Sidheekh, S., Tenali, P., Mathur, S., Blasch, E., Kersting, K., Natarajan, S.: Credibility-aware multimodal fusion using probabilistic circuits. In: Proc. AISTATS, vol. 258, pp. 2305–2313 (2025)
[9] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. ICML, vol. 139, pp. 8748–8763 (2021)
[10] Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: CLAP: Learning audio concepts from natural language supervision. In: Proc. IEEE ICASSP, pp. 1–5 (2023)
[11] Gong, Y., Chung, Y.-A., Glass, J.: AST: Audio spectrogram transformer. In: Proc. Interspeech, pp. 571–575 (2021)
[12] Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: Proc. IEEE ICASSP, pp. 776–780 (2017)
[13] Wang, S., Mesaros, A., Heittola, T., Virtanen, T.: Audio-visual scene classification: Analysis of dcase 2021 challenge submissions. In: Proc. DCASE Workshop, pp. 45–49 (2021)
[14] Okazaki, S., Kong, Q., Yoshinaga, T.: A multi-modal fusion approach for audio-visual scene classification enhanced by CLIP variants. In: Proc. DCASE Workshop, pp. 95–99 (2021)
[15] Zhou, L., Zhou, Y., Qi, X., Hu, J., Lam, T.L., Xu, Y.: Feature pyramid attention network for audio-visual scene classification. CAAI Trans. Intell. Technol. **10**, 359–374 (2025)
[16] Yin, M., Liang, T., Zeng, X., Shi, D.: Multimodal multi-scale temporal enhancement network for audio-visual scene classification. IEEE Trans. Audio Speech Lang. Process. **34**, 1944–1951 (2026)
[17] Tonami, N., Imoto, K., Yamanishi, R., Yamashita, Y.: Joint analysis of sound events and acoustic scenes using multitask learning. IEICE Trans. Inf. Syst. **E104.D**, 294–301 (2021)
[18] Hou, Y., Kang, B., Van Hauwermeiren, W., Botteldooren, D.: Relation-guided acoustic scene classification aided with event embeddings. In: Proc. IJCNN, pp. 1–8 (2022)
[19] Hou, Y., Song, S., Yu, C., Wang, W., Botteldooren, D.: Audio event-relational graph representation learning for acoustic scene classification. IEEE Signal Process. Lett. **30**, 1382–1386 (2023)
[20] Hou, Y., Kang, B., Botteldooren, D.: Audio-visual scene classification via contrastive event-object alignment and semantic-based fusion. In: Proc. IEEE MMSP (2022)
[21] Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 2880–2894 (2020)