GeoTAP: Resolving Video Geolocation Ambiguity via Temporal Audio Prototypes

Anonymous authors
Paper under double-blind review

Abstract

Determining where a video was filmed is a fundamental challenge in spatial intelligence and computer vision. While modern visual foundation models excel at geolocating images, adapting them to video streams introduces two core failure modes: spatial aggregation noise, where naïve per-frame averaging on the sphere degrades accuracy below a single static frame, and visual candidate ambiguity, where visually interchangeable scenes leave top geographic predictions tied with high confidence. To address both challenges, we propose **GeoTAP**, an encoder-agnostic framework that formulates video geolocation as hypothesis disambiguation within a unified Bayesian pipeline. GeoTAP extracts spatial candidate hypotheses via geodesic clustering and spherical geometric-median filtering to eliminate aggregation noise, then disambiguates candidate hypotheses against temporal audio patterns using learnable acoustic prototypes in our TAP (Temporal Audio Prototypes) module. Visual prior entropy naturally regulates audio influence during Bayesian posterior selection. Evaluated on our new **AVG** benchmark of $20{,}538$ continuous video clips with diegetic audio across 907 sites, GeoTAP raises city accuracy from $27.9\,\%$ to $34.2\,\%$, outperforming state-of-the-art visual and multimodal baselines without backbone retraining.

1 Introduction

Determining where a video was filmed is a fundamental challenge at the intersection of computer vision, multimedia analysis, and spatial intelligence. Accurate video geolocation supports real-world applications, including open-source intelligence, robotic navigation, and media verification. While modern visual foundation models geolocate single static images with impressive accuracy, real-world video streams capture continuous spatial-temporal trajectories and rich multimodal signals that static snapshots lack.

However, adapting image-level geolocators to video streams reveals two core failure modes. First, evaluating frames independently introduces spatial aggregation noise: per-frame predictions exhibit multi-peak spatial variance across a video clip. Because geographic space is non-convex, naïve coordinate averaging destructively interpolates between distant geographic peaks into ocean bodies or unpopulated regions, degrading city-level accuracy to 10.0% ($-17.9\,\%$ below a single frame).

Second, even when frame predictions are grouped into valid candidate locations, visual features suffer from candidate ambiguity across similar-looking environments. For example, on a walking tour near the University of San Francisco (Fig. 1), generic street grids cause visual geolocation models to split confidence between Abidjan, Côte d'Ivoire (23.2%) and Berkeley, USA (22.9%). While vision isolates ground truth within top-K candidates, visual features alone cannot break high-confidence ties. Synchronized audio captures distinct environmental signatures (e.g., transit chimes, acoustic ambience) providing orthogonal constraints. However, naïve multimodal fusion fails because environmental soundscapes are temporally sparse and globally ambiguous in isolation. Rather than estimating global location directly from audio (answering not "where does this sound originate?" but "which candidate hypothesis does it support?"), we use audio to evaluate and disambiguate candidate hypotheses generated by the visual prior.

To resolve these challenges, we propose **GeoTAP** (Fig. 2), an encoder-agnostic framework that formulates video geolocation as candidate-conditioned Bayesian hypothesis disambiguation. To eliminate spatial aggregation noise, GeoTAP constructs a visual spatial prior by applying geodesic clus-

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.

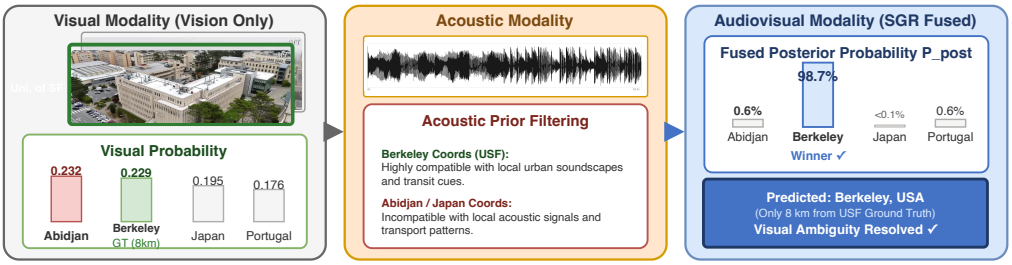

Figure 1: **Video geolocation as hypothesis disambiguation.** On a walking tour near the University of San Francisco (left), visual models split confidence between Abidjan, Côte d'Ivoire (23.2%) and Berkeley, USA (22.9%) due to generic street grids. Synchronized audio containing local transit cues (middle) filters out incompatible candidates. GeoTAP outputs the correct prediction near Berkeley (right, 98.7% posterior), resolving the visual tie.

tering and spherical geometric median estimation over per-frame predictions, identifying K compact candidate hypotheses corresponding to dominant geographic clusters on S^2. To resolve visual ambiguity, our **Temporal Audio Prototypes (TAP)** module partitions continuous audio into temporal windows, projects them into learnable acoustic prototypes, and pools compatibility scores via temporal saliency attention. The visual prior and acoustic scores combine through **entropy-calibrated Bayesian posterior selection**: visual prior entropy $\mathcal{H}(\mathbf{p}_v)$ dynamically balances modalities, keeping acoustic updates conservative when vision is confident while allowing sound to dominate during visual ties. All foundation encoders remain frozen; GeoTAP trains only a lightweight prototype adapter ($\sim$5 M parameters), avoiding costly end-to-end finetuning.

A key obstacle in video geolocation has been the lack of dedicated multimodal benchmarks containing paired video and synchronized audio. Existing benchmarks are strictly unimodal, focusing either on visual video sequences without audio (e.g., CityGuessr68k) or ambient audio without paired video (e.g., iNatSounds). To address this benchmark gap, we introduce the **AVG (Audiovisual Geolocation)** dataset, comprising 20,538 continuous diegetic video clips across 907 geographic sites. AVG applies diegetic audio filtering to remove non-diegetic background music/dubbing and enforces strict geographic clearance guards between splits to prevent spatial data leakage.

Extensive experiments demonstrate that GeoTAP achieves state-of-the-art (SoTA) results on AVG across diverse visual backbones. Starting from a visual baseline of 27.9% city accuracy, GeoTAP boosts city accuracy to 34.2% and country accuracy to 69.4% (8.2 % over GLOBE), while reducing median error from 462 km to 308 km. GeoTAP outperforms multimodal baselines and MLLM reasoning agents (*e.g.*, GeoAgent, GLOBE), while exhibiting strong zero-shot generalization to unseen external datasets (CityGuessr68k).

Contributions.

- We analyze video geolocation failure modes, demonstrating that naïve multi-frame vector averaging suffers from catastrophic spatial aggregation noise, and formulate video geolocation as candidate-conditioned Bayesian hypothesis disambiguation.
- We introduce GeoTAP, featuring prototype-based Temporal Audio Prototypes (TAP) and entropy-calibrated Bayesian posterior selection to break visual candidate ties across frozen visual encoders, backed by spatial median candidate generation.
- We release AVG, a novel dataset of 20,538 continuous video clips across 907 sites with diegetic audio filtering and strict spatial clearance guards for multimodal geolocation research.
- We demonstrate SoTA geolocation performance on AVG across diverse visual backbones and showcase zero-shot transferability to unseen datasets (CityGuessr, iNatSounds).

2 Related Work

Visual Geolocation. Visual geolocation has progressed through regression Weyand et al. (2016); Seo et al. (2018), contrastive retrieval Vivanco Cepeda et al. (2023); Yang et al. (2021); Haas et al.

(2024); Dhakal et al. (2025), coordinate diffusion Wang et al. (2025); Dufour et al. (2025), sequence matching Kulkarni et al. (2026); Shatwell et al. (2026), and spatial MLLMs Li et al. (2025); Jia et al. (2025); Jin et al. (2026); Li et al. (2024). However, applying static geolocators to continuous video streams introduces two key failure modes: (i) per-frame predictions suffer from spatial aggregation noise when averaged across non-convex geographic space, degrading accuracy below single-frame baselines; and (ii) visual features alone produce high-entropy candidate distributions over visually interchangeable environments. GeoTAP resolves both issues by combining Weiszfeld spherical geometric median estimation with candidate-conditioned acoustic evidence.

Audio Geolocation. Audio-based geolocation methods Kumar et al. (2017); Pokorny et al. (2019); Chasmai et al. (2025); Khanal et al. (2023); Sastry et al. (2025) map acoustic signals to geographic coordinates without visual input. While effective for biophony, ambient urban soundscapes are globally ambiguous in isolation (5.4% country accuracy on AVG). Furthermore, naïve late fusion or score-averaging of audio-visual features lacks temporal windowing and candidate structure. GeoTAP shows that environmental audio is most discriminative when evaluated relatively, matching temporal acoustic patterns against candidate hypotheses generated by the visual prior.

Geolocation Benchmarks. Existing visual benchmarks (e.g., CityGuessr Kulkarni et al. (2024); Astruc et al. (2024); Vyas et al. (2022); Kristoffersen et al. (2026); Liang et al. (2026)) scale video geolocation but lack synchronized audio. Web video repositories Thomee et al. (2016) contain non-geolocatable videos Astruc et al. (2024), whereas audio-only benchmarks Chasmai et al. (2025) lack paired video streams. AVG bridges this gap as a dedicated benchmark featuring paired video and diegetic audio across 907 sites, enforced with multi-stage audio filtering and strict $50\,\mathrm{km}$ spatial clearance guards to prevent spatial data leakage.

3 METHOD

3.1 PROBLEM FORMULATION AND VISUAL CANDIDATE GENERATION

Let a video clip $\mathbf{V}$ consist of T sampled frames $\{v_1, \ldots, v_T\}$ accompanied by a synchronized audio stream $\mathbf{A}$. Each frame v_k is mapped by a frozen visual backbone to a 3D unit direction vector $\mathbf{m}_k = [\cos\phi_k \cos\lambda_k, \cos\phi_k \sin\lambda_k, \sin\phi_k]^\top \in S^2$ representing latitude ϕ_k and longitude λ_k, alongside a visual gallery embedding $\mathbf{e}_k \in \mathbb{R}^{d_e}$. The goal of video geolocation is to estimate the ground-truth 3D geographic direction $\mathbf{y}^* \in S^2$ corresponding to the true filming location.

When adapting image-level foundation models to video streams, two core failure modes emerge:

1. Spatial aggregation noise: Because image-level foundation models are trained on static snapshots without inter-frame consistency constraints, per-frame visual predictions exhibit multi-peak spatial dispersion across a video clip. Because geographic candidate distributions are non-convex, taking the arithmetic or spherical mean across distinct spatial clusters destructively interpolates coordinates into unpopulated intermediate regions (e.g., oceans or wrong landmasses), degrading city accuracy to 10.0% (Sec. 5.3).
2. Visual candidate ambiguity: Even when predictions are grouped into valid candidate locations, visual features alone frequently yield high-entropy candidate distributions over visually interchangeable environments.

GeoTAP resolves both challenges by formulating video geolocation as candidate-conditioned Bayesian hypothesis disambiguation (Fig. 2). Operating over frozen visual backbones (e.g., GeoCLIP Vivanco Cepeda et al. (2023), GLOBE Li et al. (2025), GeoAgent Jin et al. (2026)), GeoTAP first extracts a set of spatially coherent candidate hypotheses to eliminate aggregation noise, then disambiguates them using temporal audio signatures.

Visual candidate prior construction. From T sampled frames, raw predictions $\{\mathbf{m}_1, \ldots, \mathbf{m}_T\} \subset S^2$ are grouped via greedy geodesic clustering with a $25\,\mathrm{km}$ radius into K candidate clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$. Each candidate center is estimated via the spherical geometric median (Weiszfeld algorithm Weiszfeld (1937)) over member frame directions:

$$\mathbf{m}_i = \text{Weiszfeld}\left(\{\mathbf{m}_k : k \in \mathcal{C}_i\}\right) \in S^2, \qquad \mathbf{e}_i = \frac{1}{|\mathcal{C}_i|} \sum_{k \in \mathcal{C}_i} \mathbf{e}_k. \tag{1}$$

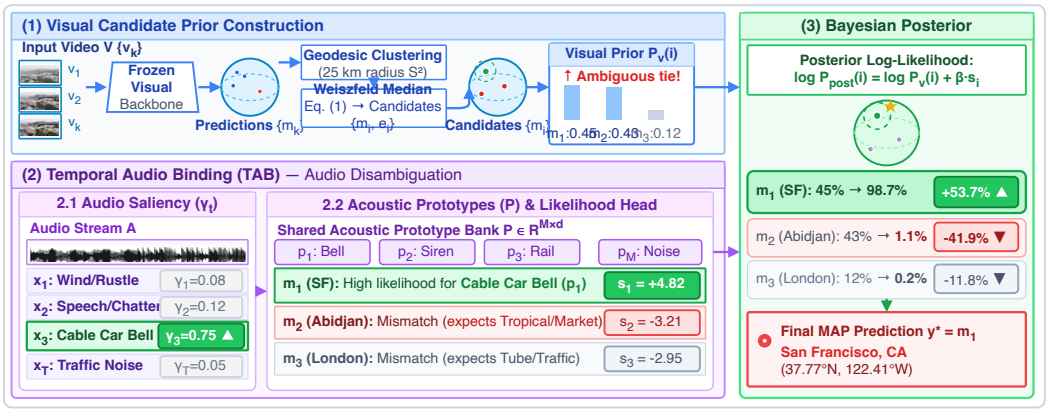

Figure 2: **Overview of the GeoTAP framework. (1) Visual Prior:** A frozen visual backbone predicts per-frame locations on S^2; greedy geodesic clustering and Weiszfeld median aggregation yield K candidate hypotheses $\{\mathbf{m}_i\}$ and visual prior $P_v(i)$. **(2) Temporal Audio Prototypes (TAP):** Audio windows $\mathbf{x}_t$ are soft-assigned to acoustic prototypes $\mathbf{P}$; each candidate predicts its expected prototype distribution $\boldsymbol{\pi}_{a,i}$; temporal saliency attention γ_t pools window log-likelihoods into candidate scores s_i. **(3) Bayesian Posterior Selection:** The posterior $\log P_{\text{post}}(i) = \log P_v(i) + \beta\, s_i$ selects the MAP candidate $\hat{\imath}$, outputting candidate center $\hat{\mathbf{y}} = \mathbf{m}_{\hat{\imath}} \in S^2$.

Unlike arithmetic vector averaging, the spherical geometric median resists up to 50% outlier frame predictions on S^2. The candidate centers define a discrete visual prior distribution $P_v(i) = \pi_i$:

$$P_v(i) = \pi_i = \frac{\exp(\tau\,|\mathcal{C}_i|/T)}{\sum_{j=1}^{K} \exp(\tau\,|\mathcal{C}_j|/T)}, \tag{2}$$

where $|\mathcal{C}_i|$ is member frame support and $\tau > 0$ is a visual temperature.

While geodesic Weiszfeld filtering isolates ground-truth regions, visual features alone leave candidate probabilities tied with high entropy ($H(P_v)/\ln K \approx 0.84$), necessitating complementary acoustic disambiguation.

3.2 Temporal Audio Prototypes (TAP)

Environmental soundscapes contain rich geographic signatures (e.g., transit bells, sirens, biophony) but are temporally sparse. Audio is partitioned into W temporal windows $\mathbf{A} = \{\mathbf{x}_1, \ldots, \mathbf{x}_W\} \subset \mathbb{R}^d$ encoded via a frozen audio encoder ($d = 512$). TAP matches these windows against candidate locations through a shared acoustic prototype vocabulary.

TAP maintains M learnable acoustic prototypes $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_M]^\top \in \mathbb{R}^{M\times d}$, each capturing a discoverable soundscape pattern (e.g., urban rail, wind, crowd noise). Each window embedding $\mathbf{x}_t \in \mathbb{R}^d$ is soft-assigned to prototypes via cosine similarity:

$$\mathbf{q}_t = \text{softmax}\left(\frac{1}{\tau_p}\hat{\mathbf{P}}\hat{\mathbf{x}}_t\right) \in \Delta^{M-1}, \tag{3}$$

where $\hat{\mathbf{P}} \in \mathbb{R}^{M\times d}$ and $\hat{\mathbf{x}}_t \in \mathbb{R}^d$ are ℓ_2-normalized prototypes and window features, and τ_p is a prototype temperature. The vector $\mathbf{q}_t$ expresses window t's acoustic content over the prototype vocabulary. For example (Fig. 1), a window with transit audio strongly activates an "urban rail" prototype, while background wind spreads mass broadly.

Geographic locations predict distinct prototype activation signatures. Each candidate hypothesis i predicts its expected prototype distribution from its location embedding $\mathbf{e}_i \in \mathbb{R}^{d_e}$:

$$\boldsymbol{\pi}_{a,i} = \text{softmax}\big(\mathbf{W}_{\text{proto}}\,\mathbf{e}_i\,/\,\tau_c\big) \in \Delta^{M-1}, \tag{4}$$

where $\mathbf{W}_{\text{proto}} \in \mathbb{R}^{M\times d_e}$ is a learnable projection and τ_c is a candidate temperature. The compatibility between window t and candidate i is the categorical log-likelihood:

$$s_{t,i} = \mathbf{q}_t^\top \log \boldsymbol{\pi}_{a,i}. \tag{5}$$

To spotlight discriminative acoustic events over non-informative background windows, TAP computes scalar temporal attention weights:

$$\gamma_t = \frac{\exp(\mathbf{w}_\gamma^\top \mathbf{x}_t)}{\sum_{t'=1}^{W} \exp(\mathbf{w}_\gamma^\top \mathbf{x}_{t'})}, \tag{6}$$

via projection vector $\mathbf{w}_\gamma \in \mathbb{R}^d$. The clip-level acoustic score for candidate i and corresponding acoustic likelihood distribution are:

$$s_i = \alpha \sum_{t=1}^{W} \gamma_t \, s_{t,i}, \qquad P_\theta(i \mid \mathbf{A}) = \frac{\exp(s_i)}{\sum_{j=1}^{K} \exp(s_j)}, \tag{7}$$

where $\alpha > 0$ is a learnable scale parameter.

Training. TAP is trained to maximize correct candidate log-likelihood while regularizing 3D direction estimates via an auxiliary direction head:

$$\mathcal{L}_{\text{TAP}}(\theta) = \mathbb{E}\big[-\log P_\theta(i^* \mid \mathbf{A}) \big] + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}}, \tag{8}$$

where i^* is the ground-truth candidate index, and $\mathcal{L}_{\text{geo}} = -\kappa_a(\hat{\mathbf{m}}_a^\top \mathbf{y}_{\text{gt}}) - \log C_3(\kappa_a)$ is a von Mises–Fisher negative log-likelihood on auxiliary direction $\hat{\mathbf{m}}_a \in S^2$ with concentration κ_a. All visual and audio encoders remain frozen; only $\mathbf{P}$, $\mathbf{W}_{\text{proto}}$, $\mathbf{w}_\gamma$, and α are fitted ($\sim$5 M parameters).

Why audio alone is not enough. While audio provides discriminative acoustic evidence, audio-only selection achieves top-1 accuracy on only 23% of clips compared to 56% for the visual prior. However, audio excels as a filter: correct candidates reach rank $\leq$2 in 56% of clips, demonstrating that sound complementarily eliminates false-positive visual candidates.

3.3 Bayesian Posterior Selection

We formulate candidate selection as discrete Bayesian Maximum A Posteriori (MAP) inference over visual spatial hypotheses $i \in \{1, \ldots, K\}$. Assuming conditional independence between acoustic evidence and camera viewing angle given candidate location i ($P(\mathbf{A} \mid i, \mathbf{V}) \approx P(\mathbf{A} \mid i)$), Bayes' rule expresses posterior candidate belief as $P(i \mid \mathbf{V}, \mathbf{A}) \propto P(i \mid \mathbf{V})\, P(\mathbf{A} \mid i)^\beta$. Taking the logarithm yields the MAP decision rule:

$$\log P_{\text{post}}(i) = \log P_v(i) + \beta \, s_i + \text{const}, \qquad \hat{\imath} = \arg\max_i \log P_{\text{post}}(i), \tag{9}$$

where $P_v(i)$ acts as the visual hypothesis prior, acoustic score $s_i = \log P_\theta(i \mid \mathbf{A})$ represents candidate-conditioned log-likelihood, and $\beta > 0$ is a likelihood temperature parameter controlling multimodal calibration. The final prediction reports the MAP candidate center $\hat{\mathbf{y}} = \mathbf{m}_{\hat{\imath}} \in S^2$.

The relative belief between any two visual candidate hypotheses i and j follows the log-odds differential:

$$\log \frac{P_{\text{post}}(i)}{P_{\text{post}}(j)} = \log \frac{P_v(i)}{P_v(j)} + \beta \underbrace{(s_i - s_j)}_{\Delta s_{ij}}. \tag{10}$$

where the acoustic score differential $\Delta s_{ij} = s_i - s_j$ adjusts visual log-odds based on relative soundscape evidence. This Bayesian formulation naturally adapts to visual confidence: when P_v is peaked, visual evidence dominates; when P_v is high-entropy, s_i acts as a decisive tie-breaker. Empirically, audio reranking improves city accuracy by 3.6 % in the highest visual entropy quintile while preserving predictions in the lowest quintile (Sec. 5.3).

Computational complexity. Inference requires $\mathcal{O}(TC_v + WC_a + KWM + K)$, where C_v and C_a are per-frame visual and per-window audio encoding costs. The entire pipeline processes a 15 s clip in 0.17 s total latency on a single GPU, running $\sim$15$\times$ faster than MLLM-based geolocators.

Table 1: Comparison with existing geolocation datasets. AVG is the first video benchmark with synchronized diegetic audio and spatially guarded location-disjoint splits.

Dataset	Modality	Clips	Locations	Diegetic Audio	Spatial Guard
CityGuessr Kulkarni et al. (2024)	Video only	68k	166 cities	×	×
YFCC100M Thomee et al. (2016)	Image+audio	100M+	varied	×	×
iNatSounds Chasmai et al. (2025)	Audio only	1M+	varied	✓	×
AVG (ours)	**Video+audio**	20,538	907	✓	50 km

4 AVG: A BENCHMARK FOR AUDIOVISUAL DISAMBIGUATION

Can environmental audio disambiguate visually similar locations? Answering this question requires a benchmark where: (1) audio is *diegetic*, physically produced by the environment; (2) locations are geographically diverse; and (3) splits prevent spatial leakage between training and evaluation. No existing benchmark satisfies all three criteria (Tab. 1). Visual benchmarks like CityGuessr Kulkarni et al. (2024) lack audio; web repositories like YFCC100M Thomee et al. (2016) contain non-geolocatable videos; and bioacoustic datasets like iNatSounds Chasmai et al. (2025) lack paired video. We introduce **AVG** (Audiovisual Geolocation) as the first video-audio benchmark designed for multimodal candidate disambiguation.

Dataset Collection Pipeline. To build AVG, we design a three-stage curation pipeline:

1. **Geographic Sampling and Video Acquisition.** Target locations are sampled from GeoNames GeoNames (2024) across six continents, stratified across major urban centers (45%), medium cities (25%), landmarks (20%), and natural sites (10%). Raw ground-level videos are retrieved from YouTube using first-person walking-tour query templates and title heuristics to ensure environmental grounding (Appendix A), and segmented into continuous 10–20 s clips.
2. **Diegetic Audio Filtering.** Uncurated web audio often decouples sound from environmental physics. To preserve genuine acoustic signatures, such as regionally distinctive transit chimes, siren patterns, and ambient biophonies, we pass candidate clips through a multi-stage audio filtering pipeline (see Appendix B). This pipeline automatically rejects: (i) non-diegetic audio overlays (background music, sound effects); (ii) speech-dominant clips, eliminating potential place-name leakage via speech recognition and forcing reliance on acoustic cues; and (iii) degraded audio (severe wind noise, clipping, or silence). This stage discards $\sim$50% of raw clips, ensuring all retained audio strictly co-refers to the filmed physical environment.
3. **Spatial Clearance Guard and Dataset Statistics.** To prevent spatial leakage between training and evaluation, every validation and test location must be at least 50 km away from the nearest training location. AVG contains 20,538 clips across 907 location-disjoint sites, split into train (16,431), validation (2053), and test (2054) clips.

5 EXPERIMENTS

5.1 SETUP

Benchmarks. *AVG* (Sec. 4) is our primary benchmark, comprising 20,538 diegetic-audio clips across 907 geographic sites with location-disjoint train, validation, and test splits (16,431 / 2,053 / 2,054). *CityGuessr* Kulkarni et al. (2024) provides street-level walking videos across 166 cities without audio; we feed a zero waveform to evaluate framework robustness under missing audio. *iNatSounds* Chasmai et al. (2025) contains geotagged natural sound recordings without video; we evaluate zero-shot audio-only transfer by constructing candidate sets via audio retrieval.

Encoders and training protocol. For visual candidate generation, we sample $T = 12$ frames per clip and encode them with a frozen visual backbone (GLOBE Li et al. (2025), GeoCLIP Vivanco Cepeda et al. (2023), or GeoAgent Jin et al. (2026)). Audio is segmented into 3 s windows at 1.5 s stride and encoded with frozen GeoCLAP Khanal et al. (2023) ($d = 512$). All visual and audio foundation backbones remain frozen throughout training; only the TAP prototype adapter and

Table 2: **Audiovisual geolocation performance on the AVG test set.** Accuracy (%) across geographic scales (City, Region, Country, Continent) and Median Distance Error (km). Best performance within each model category is in **bold**. GeoTAP consistently improves geolocation accuracy across diverse frozen visual foundation models. Small teal numbers (+·) show improvement over the corresponding visual baseline.

Method	City	Region	Country	Continent	Median Dist. (km)
Visual-only baselines (video frames)					
GLOBE	27.9	35.5	61.2	80.9	462
GeoCLIP	4.1	19.3	46.1	75.1	852
RFM S^2	14.2	28.1	55.3	78.9	589
GeoReasoner	12.4	25.9	41.9	72.3	1104
GeoAgent	21.1	31.0	58.8	81.5	507
Audio-only baselines (clip audio)					
GeoCLAP	0.0	0.8	5.4	18.7	7545
TaxaBind	0.0	0.2	3.0	11.8	8924
Ours: + TAP (audio disambiguation)					
GeoCLIP + Ours	4.4 (+0.3)	21.7 (+2.4)	51.3 (+5.2)	80.3 (+5.2)	718 (-134)
GeoAgent + Ours	26.9 (+5.8)	36.5 (+5.5)	67.1 (+8.3)	85.0 (+3.5)	373 (-134)
GLOBE + Ours	34.2 (+6.3)	43.4 (+7.9)	69.4 (+8.2)	86.2 (+5.3)	308 (-154)

saliency projection are fitted ($\sim$5.1 M parameters). We evaluate under two protocols: *In-Distribution* (trained and evaluated on AVG) and *Out-of-Distribution (OOD)* (trained on AVG and evaluated zero-shot on CityGuessr or iNatSounds).

Baselines. We compare against SoTA geolocators: *Visual-only baselines* include deep visual retrieval model GeoCLIP Vivanco Cepeda et al. (2023), coordinate diffusion model RFM S^2 Dufour et al. (2025), and MLLM-based reasoning frameworks GeoReasoner Li et al. (2024), GLOBE Li et al. (2025), and GeoAgent Jin et al. (2026). *Audio-only baselines* include GeoCLAP Khanal et al. (2023) and TaxaBind Sastry et al. (2025). Standard multimodal fusion strategies (e.g., late score fusion, coordinate interpolation) are systematically analyzed in Sec. 5.3 and Tab. 4(b). All video baselines process the identical 12-frame sequence per clip.

5.2 Results

Quantitative results and baseline comparisons. Table 2 presents quantitative evaluation on AVG test set. GeoTAP paired with the GLOBE visual backbone sets a new SoTA across all metrics, achieving 34.2% city accuracy, 43.4% region accuracy, 69.4% country accuracy, 86.2% continent accuracy, and reducing median error to 308 km (6.3 % city and 8.2 % country accuracy over the single-frame baseline). Comparing across baseline paradigms highlights three key insights: (i) *MLLMs and reasoning agents* (GeoReasoner: 41.9% country, GeoAgent: 58.8% country) suffer from frame-level spatial variance across video streams without S^2 candidate hypothesis clustering; (ii) *Standalone audio models* (GeoCLAP: 5.4% country) lack spatial density for direct global coordinate regression; and (iii) *Naïve multimodal fusion* (e.g., late score fusion: 59.7% country) degrades visual priors by injecting uncalibrated audio noise, whereas GeoTAP's Bayesian posterior selection outperforms late fusion by 9.7 % in country accuracy and 194 km median error reduction.

Encoder-agnostic adaptability. Evaluating TAP across diverse visual backbones (Tab. 2) demonstrates consistent plug-and-play gains: country accuracy increases by 5.2 % with GeoCLIP and 8.3 % with GeoAgent (cutting median error by 134 km). This confirms GeoTAP as a flexible, model-agnostic disambiguation head across retrieval models, foundation encoders, and MLLMs.

Generalization to Unseen Data (CityGuessr). On the video-only CityGuessr dataset (166 cities, no audio), we evaluate zero-shot transfer across four visual backbone paradigms by feeding a constant zero waveform (Tab. 3). Among single-frame visual baselines, MLLM-based geolocators substantially outperform retrieval: GLOBE achieves 78.8% country accuracy, followed by GeoAgent (75.9%), GeoReasoner (63.0%), and GeoCLIP (61.0%), consistent with the AVG ranking (Tab. 2).

Table 3: Zero-shot video geolocation on CityGuessr68k (no audio). Visual-only baselines use a single mid-frame; GeoTAP aggregates all frames and receives zero waveforms, so acoustic log-odds vanish (Eq. (10)) and posterior selection defaults to the visual prior. Small teal numbers (+·) show zero-shot improvement over the single-frame baseline.

Method	City	Region	Country	Continent	Median Dist. (km)
Visual-only baselines (single frame)					
GeoCLIP	10.4	31.1	61.0	83.9	474
GeoReasoner	39.1	42.6	63.0	80.0	393
GeoAgent	53.0	58.9	75.9	88.9	8
GLOBE	57.7	63.0	78.8	91.0	<1
+ GeoTAP (OOD, zero waveform)					
GeoCLIP + Ours	11.2 (+0.8)	36.2 (+5.1)	68.8 (+7.8)	88.8 (+4.9)	361 (-113)
GeoReasoner + Ours	42.3 (+3.2)	46.2 (+3.6)	69.1 (+6.1)	88.4 (+8.4)	297 (-96)
GeoAgent + Ours	56.2 (+3.2)	62.9 (+4.0)	79.5 (+3.6)	91.9 (+3.0)	5 (-3)
GLOBE + Ours	61.0 (+3.3)	66.5 (+3.5)	81.1 (+2.3)	92.5 (+1.5)	<1

GeoTAP with GLOBE sets the highest OOD accuracy (81.1% country, 92.5% continent), improving over the single-frame GLOBE baseline by 2.3 and 1.5 percentage points, respectively. Under a zero waveform ($\mathbf{x}_t = \mathbf{0}$), acoustic log-odds vanish (Eq. (10)) and posterior selection defaults to the visual prior, so these gains stem entirely from multi-frame aggregation—confirming both the benefit of temporal evidence pooling and the graceful degradation of TAP under missing audio.

5.3 Analysis & Ablations

Standard frame aggregation methods. Table 4(a) evaluates multi-frame aggregation strategies across T=12 sampled video frames. Mean-based pooling (coordinate: 10.1% city, spherical: 10.0%) collapses accuracy from the single-frame baseline (27.9%) because averaging non-convex, multi-peak geographic predictions destructively interpolates coordinates into unpopulated regions. Robust estimators like coordinate median (28.4%) and spherical median (31.2%) recover performance, confirming outlier sensitivity as the primary failure mode. GeoTAP achieves 34.2% city accuracy by grouping predictions within a 25 km radius.

Standard audiovisual fusion methods. Table 4(b) evaluates standard fusion baselines combining visual predictions (GLOBE: 27.9% city) with independent GeoCLAP audio retrieval. All standard fusion strategies degrade the visual baseline: coordinate averaging (0.1%), α-weighted interpolation (5.5%), late score fusion (22.5%), and gated switching (20.6%). Because raw audio predictions inject spatial noise into visual priors, direct mixing fails. GeoTAP operates on discrete candidate hypotheses via candidate-conditioned Bayesian posterior selection, achieving 34.2% city accuracy.

Component ablation and sensitivity. Table 5 ablates GeoTAP's components and prototype count M. Acoustic prototype scores s_i alone achieve 61.6% country accuracy, confirming strong candidate-conditioned discriminability over restricted candidate sets. Combining visual spatial priors with TAP scores via Bayesian MAP posterior selection (Eq. (9)) yields optimal accuracy (34.2% city, 69.4% country, 308 km median error). Removing learned temporal saliency (replacing γ_t with uniform weights) reduces country accuracy by 0.9 %, confirming that learned attention over acoustic windows contributes to geographic discrimination. Varying the prototype count $M \in \{8, 16, 32, 64\}$ shows consistent performance across all values, with M=32 achieving the best balance; even M=8 retains 68.4% country accuracy, demonstrating that the method is not sensitive to this hyperparameter. As analyzed in Sec. C.2, audio reranking delivers the largest gains (3.6 % city acc) when visual priors are ambiguous (top entropy quintile), while leaving confident visual predictions untouched.

Qualitative candidate disambiguation. As illustrated in Fig. 1, visual model GLOBE splits confidence between Abidjan (23.2%) and Berkeley (22.9%) due to generic street poles. Synchronized audio containing transit chimes activates Bay Area prototypes, shifting the posterior to 98.7% for Berkeley. In both urban and coastal examples, temporal saliency attention γ_t (Eq. (6)) concentrates mass on informative soundscape windows while suppressing background noise.

Table 4: Standard frame aggregation and audiovisual fusion methods on AVG test. (a) Mean-based pooling collapses city accuracy; robust estimators recover. (b) All naïve fusion strategies degrade the visual baseline. Parenthetical numbers show performance gain (+·) or drop (-·) relative to the single-frame GLOBE baseline.

Method	Accuracy (%)				Median Dist. (km)
	City	Region	Country	Continent	
(a) Frame aggregation (vision-only, GLOBE × 12 frames)					
GLOBE	27.9	35.5	61.2	80.9	462
Coordinate mean	10.1 (-17.8)	26.6 (-8.9)	56.9 (-4.3)	81.8 (+0.9)	552 (+90)
Spherical mean	10.0 (-17.9)	26.8 (-8.7)	60.5 (-0.7)	84.8 (+3.9)	513 (+51)
Coordinate median	28.4 (+0.5)	40.7 (+5.2)	67.9 (+6.7)	85.3 (+4.4)	343 (-119)
Spherical median	31.2 (+3.3)	42.6 (+7.1)	69.0 (+7.8)	86.6 (+5.7)	328 (-134)
(b) Audiovisual fusion (GLOBE mid + GeoCLAP audio)					
GLOBE	27.9	35.5	61.2	80.9	462
Coordinate mean	0.1 (-27.8)	1.1 (-34.4)	9.8 (-51.4)	31.7 (-49.2)	4011 (+3549)
α-weighted interp. (α=0.99)	5.5 (-22.4)	36.0 (+0.5)	61.7 (+0.5)	80.9 (+0.0)	455 (-7)
Late score fusion (s_v+s_a)	22.5 (-5.4)	30.4 (-5.1)	59.7 (-1.5)	81.4 (+0.5)	502 (+40)
Gated fusion (top-25% audio)	20.6 (-7.3)	26.6 (-8.9)	46.6 (-14.6)	66.2 (-14.7)	951 (+489)
GeoTAP (Ours)	34.2 (+6.3)	43.4 (+7.9)	69.4 (+8.2)	86.2 (+5.3)	308 (-154)

Table 5: Component ablation and prototype count sensitivity on AVG test. Top: modality decomposition. Middle: architectural ablation. Bottom: prototype count M.

Configuration	Accuracy (%)				Median Dist. (km)
	City	Region	Country	Continent	
Modality decomposition					
GLOBE	27.9	35.5	61.2	80.9	330
TAP scores only (no prior)	23.6	33.7	61.6	82.2	463
Component ablation					
w/o saliency (uniform γ_t)	34.2	43.6	68.5	85.3	303
Prototype count sensitivity					
M=8	33.9	42.9	68.4	85.5	326
M=16	33.9	43.2	68.8	85.5	312
M=64	33.8	43.0	68.7	85.4	321
GeoTAP (M=32)	34.2	43.4	69.4	86.2	308

Computational efficiency. By keeping foundation backbones frozen, the trainable TAP head adds only $\sim$5.1 M parameters (1.2% parameter overhead) and 0.015 s latency (9% runtime share, total clip processing time 0.17 s per GPU, see Sec. C.5). Consequently, GeoTAP runs $\sim$15$\times$ faster than MLLM-based geolocators while achieving higher accuracy.

6 Conclusion

Video geolocation with visual foundation models faces two challenges that single-image methods leave unresolved: aggregating noisy per-frame predictions into reliable candidates, and disambiguating among them when vision alone cannot commit. GeoTAP addresses both in a unified framework: geodesic clustering with Weiszfeld estimation constructs a robust visual prior, and prototype-based temporal audio prototyping provides the complementary signal needed to break visual ties. The Bayesian posterior naturally adapts: audio dominates only where the visual prior is ambiguous, preserving confident predictions elsewhere. Evaluated on AVG, a benchmark purpose-built for diegetic audiovisual disambiguation, the framework improves geolocation across contrastive and vision–language backbones with a lightweight, encoder-agnostic head.

AI USE STATEMENT

In this work, generative AI tools were utilized solely for text polishing, formatting, and copyediting assistance. All research ideas, mathematical formulations, experimental pipelines, dataset curation protocols, and scientific conclusions were developed entirely by the authors.

ETHICS STATEMENT

GeoTAP processes public media streams to perform spatial geolocation. The AVG benchmark consists exclusively of publicly available walking-tour video clips containing no personally identifiable information (PII) or private data. All dataset metadata and geographic coordinates adhere to public licensing terms (GeoNames, YouTube API terms of service).

REPRODUCIBILITY STATEMENT

To ensure complete reproducibility, all model architectures, hyperparameter settings, visual candidate generation algorithms, and training protocols are detailed in Sec. 3, Sec. 5.1, and the Supplementary Material. Source code, dataset manifests, and pre-trained prototype weights will be made publicly available.

REFERENCES

Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, et al. Openstreetview-5m: The many roads to global visual geolocation. In *CVPR*, 2024.

Mustafa Chasmai, Wuao Liu, Subhransu Maji, and Grant Van Horn. Audio geolocation: A natural sounds benchmark. *arXiv preprint arXiv:2505.18726*, 2025.

Aayush Dhakal, Srikumar Sastry, Subash Khanal, Adeel Ahmad, Eric Xing, and Nathan Jacobs. Range: Retrieval augmented neural fields for multi-resolution geo-embeddings. In *CVPR*, 2025.

Nicolas Dufour, Vicky Kalogeiton, David Picard, and Loic Landrieu. Around the world in 80 timesteps: A generative approach to global visual geolocation. In *CVPR*, 2025.

GeoNames. GeoNames geographic database. `https://www.geonames.org/`, 2024. Accessed 2024.

Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *CVPR*, 2024.

Pengyue Jia, Seongheon Park, Song Gao, Xiangyu Zhao, and Sharon Li. Georanker: Distance-aware ranking for worldwide image geolocalization. In *NeurIPS*, 2025.

Modi Jin, Yiming Zhang, Bo-Yuan Sun, Dingwen Zhang, Ming-Ming Cheng, and Qibin Hou. Geoagent: Learning to geolocate everywhere with reinforced geographic characteristics. In *CVPR*, 2026.

Subash Khanal, Srikumar Sastry, Aayush Dhakal, and Nathan Jacobs. Learning tri-modal embeddings for zero-shot soundscape mapping. In *BMVC*, 2023.

Oskar Kristoffersen, Alba Reinders Sánchez, Morten Rieger Hannemose, Anders Bjorholm Dahl, and Dim P Papadopoulos. Mmlandmarks: a cross-view instance-level benchmark for geo-spatial understanding. In *CVPR*, 2026.

Parth Parag Kulkarni, Gaurav Kumar Nayak, and Mubarak Shah. Cityguessr: City-level video geolocalization on a global scale. In *ECCV*, 2024.

Parth Parag Kulkarni, Gaurav Kumar Nayak, and Mubarak Shah. Vidtag: Temporally aligned video to gps geolocalization with denoising sequence prediction at a global scale. In *CVPR*, 2026.

Anurag Kumar, Benjamin Elizalde, and Bhiksha Raj. Audio content based geotagging in multimedia. In *Proc. Interspeech 2017*, 2017.

Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *ICML*, 2024.

Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. In *NeurIPS*, 2025.

Xiao Liang, Huaizhi Tang, Feiyang Zhang, Shiji Yuan, Chun Hu, Dezhi Zheng, and Kang Ma. Unigeors: A unified benchmark for tri-view geo-localization. In *CVPR*, 2026.

Florian B Pokorny, Moritz Fišer, Franz Graf, Peter B Marschik, and Björn W Schuller. Sound and the city: Current perspectives on acoustic geo-sensing in urban environment. *Acta Acustica United with Acustica*, 2019.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1765–1774. IEEE, 2025.

Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *ECCV*, 2018.

David G Shatwell, Sirnam Swetha, and Mubarak Shah. Tiger: A unified framework for time, images and geo-location retrieval. In *CVPR*, 2026.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *NeurIPS*, 2023.

Shruti Vyas, Chen Chen, and Mubarak Shah. Gama: Cross-view video geo-localization. In *ECCV*, 2022.

Zhangyu Wang, Zeping Liu, Jielu Zhang, Zhongliang Zhou, Qian Cao, Nemin Wu, Lan Mu, Yang Song, Yiqun Xie, Ni Lao, and Gengchen Mai. Locdiff: Identifying locations on earth by diffusing in the hilbert space. In *NeurIPS*, 2025.

Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386, 1937.

Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *ECCV*, 2016.

Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. In *NeurIPS*, 2021.

A AVG DATA COLLECTION

This supplement documents the full AVG construction pipeline described in the main paper. All collection stages are **resumable**: manifests and per-video caches record progress so new locations and videos can be added incrementally.

A.1 PIPELINE OVERVIEW

Collection proceeds in eight stages:

1. **Location sampling**: draw stratified sites from GeoNames GeoNames (2024).
2. **Video discovery**: query YouTube with category-specific search templates.
3. **Metadata curation**: normalize location IDs, deduplicate, and apply secondary title filters.
4. **Download**: acquire MP4 files with synchronized audio (720p cap).
5. **Temporal segmentation**: propose 10–20 s clip windows via shot detection.
6. **Quality filtering**: score audio and visual diegetic alignment.
7. **Coverage-first selection**: choose export clips under per-location and per-video caps.
8. **Export**: re-encode to H.264/AAC and write training manifests.

A.2 GEOGRAPHIC LOCATION SAMPLING

Source. Locations are drawn from the GeoNames `allCountries` gazetteer. Each record includes a GeoNames ID, WGS84 coordinates, country code, feature class/code, and a category label for stratified sampling.

Target distribution. We target approximately 1000 locations with at least one usable video:

Table 6: Location category quotas.

Category	Share	Sampling criterion
Major cities	45%	Populated places with population $\geq$ 1M
Medium cities	25%	Populated places, 50k–1M population
Landmarks	20%	Museums, towers, palaces, universities, stadiums with $\geq$ 5 alternate names
Natural	10%	Terrain/landscape features with $\geq$ 8 alternate names

Generic landmark names (*e.g.*, "church", "chapel") are excluded. The natural pool is capped to the 10k most prominent features by alternate-name count.

Homonym control. Single-token place names with weak population or prominence signals are flagged as homonym-risk (*e.g.*, "Springfield", "Victoria"). Major cities, well-known landmarks, and prominent natural sites are exempt.

Randomization. Order within each category is shuffled with fixed seed 42 for reproducibility. The collector iterates categories until per-category targets are met or the candidate pool is exhausted.

A.3 YOUTUBE VIDEO DISCOVERY

Search templates. Queries are designed to surface first-person ambient footage rather than edited productions. Examples include:

- Major cities: "*name, country* 4K walking tour", "*name* city walk", "*name* street ambience"
- Medium cities: "*name* walking tour", "*name* city walk"
- Landmarks: "*name* tour", "*name* visit" (stadium/museum-specific variants for those types)
- Natural: "*name* hiking walking tour", "*name* nature walk", "*name* trail walk"

Up to five search results per query are retrieved; collection stops at five videos per location.

Title inclusion rules. A video title must satisfy *all* of the following:

- Contains at least one ambience/walk keyword (*e.g.*, *walk, walking, tour, ambience, ambient, soundscape, 4k, stroll, visit, explore*; natural sites also accept *hike, hiking, trail, nature, park, scenic, drone, national*).
- Contains a location name token from the target place (normalized name or GeoNames alternates).
- Does not match reject keywords (music videos, gameplay, ASMR, podcasts, documentaries, official MVs, *etc.*).
- Does not match promotional keywords (ads, sponsored content, giveaways).
- Does not indicate a country/location conflict (*e.g.*, a Somalia query returning clearly Kenyan place names without Somalia context).

Rejected titles are logged with reason codes (`reject_keyword`, `promo`, `wrong_place`, `missing_required_keyword`, `location_mismatch`).

Metadata fields. Each accepted record stores video URL and title, location ID/name/coordinates, category, country, the search query that found it, and a coarse `capture_environment` label inferred from title keywords (`outdoor`, `indoor`, or `mixed`).

Query throttling. YouTube searches use randomized delays between queries and periodic long pauses. Failed or rate-limited queries are retried once after a cooldown.

A.4 METADATA CURATION

Raw discovery output undergoes a second pass before download:

1. **Location ID normalization**: map legacy IDs to {`category`}_{`geonameid`} via GeoNames coordinate and name lookup.
2. **Additional title filtering**: remove weak titles (*e.g.*, wastewater facility, rugby, virtual tour, news, documentary, superstore).
3. **Hard reject keywords**: *music video, official video, lofi, asmr, gameplay, soundtrack*.
4. **Deduplication**: one entry per video URL.

Final collection statistics. The following counts reflect the published AVG benchmark dataset.

Table 7: AVG dataset summary statistics.

Statistic	Count
Curated videos	3,802
Unique location sites	907
Successfully downloaded videos	1,378
Exported diegetic-audio clips	20,538
Train split clips	16,431
Validation split clips (location-disjoint, $\geq 50\,\mathrm{km}$ guard)	2,053
Test split clips (location-disjoint, $\geq 50\,\mathrm{km}$ guard)	2,054

Category breakdown of curated videos: major cities 1,972; medium cities 924; landmarks 609; natural 297. Title-inferred capture environment: $\sim$74% mixed, $\sim$20% outdoor, $\sim$6% indoor.

A.5 VIDEO DOWNLOAD

Videos are downloaded as MP4 with synchronized audio, capped at 720p to balance quality and storage. Download order is **diversity-first**: for each location with zero downloads, fetch one video before taking a second video at any location that already has one. Each attempt is logged with status, local path, resolution, and error message; completed files are skipped on resume.

B SEGMENTATION, FILTERING, AND SELECTION

B.1 TEMPORAL SEGMENTATION

Downloaded videos are split into candidate clips of **10–20 s** (15 s target). Shot boundaries are detected with content-based scene detection; shots shorter than 5 s are merged with neighbors. Boundaries are cached per video so later pipeline runs reuse them without re-decoding.

Windowing rules. For each detected shot:

- 10–20 s: single centered window.
- 20–40 s: split into $\sim$15 s segments.
- $>$40 s: overlapping $\sim$15 s windows with $\sim$10 s stride.

Up to 20 candidate windows per video are retained. When shot detection yields fewer than 20 candidates (common for long continuous takes), uniformly spaced windows (`uniform_fill`) are added across the timeline. Roughly 91% of candidates are scene-based; 9% are uniform-fill.

Candidates are skipped at probe time if the source lacks audio or video, duration is under 15 s, or height is below 360p.

B.2 CLIP QUALITY FILTERING

Each candidate is scored on audio and visual signals. Clips failing any threshold are marked `keep: false` with a `reject_reason`. The goal is **diegetic audiovisual correspondence**: the model should hear the environment it sees.

Table 8: Audio filter thresholds.

Signal	Captures	Reject if
Speech score	Energy/modulation in 300–3400 Hz band	$> 0.45 \rightarrow$ `speech`
Music score	Harmonic peaks, tonality, periodicity	$> 0.50 \rightarrow$ `music`
Wind score	High-frequency energy + clipping	$> 0.65 \rightarrow$ `wind`
Silence ratio	Fraction of near-silent 100 ms frames	$> 55\% \rightarrow$ `silence`
Clip ratio	Fraction of saturated samples	$> 8\% \rightarrow$ `audio_clipping`

Audio criteria.

Visual criteria. Three frames per candidate (at 25%, 50%, 75% of clip duration) are analyzed:

- Mean luminance $< 12 \rightarrow$ `black_frame`.
- Laplacian variance (sharpness) $< 35 \rightarrow$ `blur`.
- Inter-frame difference $< 0.004 \rightarrow$ `static`.

Filter outcomes (interim snapshot). Of $\sim$25,900 candidates, 13,297 ($\sim$50%) are kept. Rejections: speech 7,778; music 4,320; silence 951; other (static, blur, wind, clipping, black) 163. Speech and background music dominate, consistent with the diegetic-audio design goal.

B.3 CLIP SELECTION AND EXPORT

Filtering yields multiple valid candidates per video; selection chooses exports under coverage constraints.

Pass 1: Geographic coverage (default).

- At most one clip per location and one clip per source video.
- Among surviving candidates from each video, keep the clip with largest **quality margin** (distance above all filter thresholds).
- Process locations in diversity-first order: prioritize places with fewest selected clips.

This yields one representative clip per location before deepening coverage.

Table 9: Compatibility scaling β under different training protocols on AVG test.

Configuration	β	Country acc. (%)
Visual prior with clip-level audio	1.47	41.0
GeoTAP, clip gate + default β	1.47	$\sim$40.5
GeoTAP, clip gate + val-tuned β	2.25	42.3
GeoTAP, TAP-retrained gate	1.85	42.3

Pass 2+: Per-location diversity (optional). When re-run with higher caps:

1. Prefer clips from videos not yet represented at that location.
2. Then allow additional scenes from the same video, maximizing temporal separation.
3. Fill each location up to the per-location and per-video caps.

The main paper targets $\sim$20 clips per location at full depth (20,000 total); the interim snapshot above reflects pass 1 only.

Exported artifacts. Each clip is stored as `clips/videos/{location_id}/{clip_id}.mp4` with `clip_id = {video_id}_{start_ms}_{end_ms}`. Training manifests record clip and video IDs, coordinates, location metadata, temporal span, filter scores, quality margin, selection pass/mode, and segment method (`scene_window` or `uniform_fill`). Clips are re-encoded to H.264 + AAC at export.

B.4 Design rationale and limitations

Why ground-level ambient YouTube videos? Public first-person ambient and walking-tour videos serve as an ideal data harvesting source: they provide naturally synchronized audio and video at scale, cover diverse geographies, and capture rich ground-level sensory detail (traffic, crowds, architecture, regional ambience) that static image datasets lack.

Why filter speech and music aggressively? Audiovisual geolocation requires associating *environmental* sounds with visual context. Narration and background music decouple audio from scene geometry and introduce label noise.

Why coverage before depth? With thousands of target locations and costly download/processing, one high-quality clip per location first ensures global span. Secondary passes add intra-location diversity without letting a few long videos dominate.

Scope and design constraints.

- **Metadata validation**: video relevance is maintained via automated title heuristics, metadata verification, and diegetic acoustic filtering across all 20,538 clips.
- **Acoustic filtering**: filter thresholds are optimized for high-signal urban and suburban acoustic environments.
- **Uniform-fill windows**: when shot detection identifies minimal visual transitions, temporal windows are evenly spaced to preserve continuous audio-visual alignment.
- **Gazetteer coordinates**: location labels correspond to standardized GeoNames centroids representing site-level geographic regions.

C Additional Experiments

C.1 Compatibility scaling

Table 9 compares compatibility coefficients when pairing TAP with gates trained under different protocols. Window-level scores require higher β than clip-level scores; a gate retrained on TAP features learns a compatible scale without manual tuning.

C.2 PERFORMANCE VS. RETRIEVAL ENTROPY

We stratify the test split by prior VLM entropy $H(\boldsymbol{\pi})/\ln K$ to evaluate how prior uncertainty affects candidate fusion. As shown in Tab. 10, the accuracy gains of temporal audio prototypes over the visual prior are heavily concentrated in high-uncertainty scenarios.

Table 10: Prior VLM entropy stratification vs. city accuracy on AVG test ($n = 2054$).

Entropy Slice	Clips (n)	Visual Prior City Acc. (%)	TAP City Acc. (%)	Δ (%)
Low entropy	1303	44.1	45.0	+0.9
High entropy (top 40%)	751	14.0	16.5	+2.5
Top quintile (bin 4)	411	11.2	14.8	+3.6
Bottom quintile (bin 0)	398	62.3	62.3	0.0

C.3 ZERO-SHOT AUDIO TRANSFER ON NATURAL SOUNDSCAPES (INATSOUNDS)

On iNatSounds (geotagged audio recordings without paired video), we evaluate whether the TAP head transfers zero-shot to natural bioacoustics. We replace the visual prior with candidate locations generated via clip-level audio retrieval, and apply the AVG-trained TAP head zero-shot. As detailed in Tab. 11, OOD GeoTAP reduces median distance error by $972\,\mathrm{km}$ compared to the GeoCLAP retrieval baseline ($8514\,\mathrm{km}$ vs $9486\,\mathrm{km}$). This confirms that the learned acoustic prototype vocabulary $\mathbf{P}$ captures domain-general acoustic-geographic patterns, such as climate-specific ambient textures and biophonies, that transfer zero-shot from urban environment recordings to wild natural soundscapes.

Table 11: Audio-only geolocation on iNatSounds.

Method	Train	Accuracy (%)			Median Dist. (km)
		Region	Country	Continent	
GeoCLAP (Audio Prior)	—	0.13	1.63	8.64	9486
GeoTAP (OOD)	AVG	0.27	2.83	13.98	8514

C.4 GLOBAL SEARCH LIMITS AND CANDIDATE RERANKING UNDER GEOCLIP 1M GALLERY

GeoCLIP relies inherently on a 1M Fibonacci sphere grid as its gallery to perform spatial coordinate mapping. Table 12 evaluates candidate reranking strategies over GeoCLIP candidates alongside global audio-only retrieval baselines. Crucially, attempting direct global retrieval from audio alone (GeoCLAP and TaxaBind) across the 1M gallery fails completely (0.00% city accuracy, $\leq 2.97\%$ country accuracy, $\sim 9000\,\mathrm{km}$ median error). This empirical collapse confirms that acoustic signatures lack the spatial resolution required for unconditioned global point regression, demonstrating why candidate-conditioned Bayesian disambiguation (visual top-K candidate proposal followed by TAP audio reranking) is essential. Under this formulation, TAP reranking consistently reduces median geodesic error by $152\,\mathrm{km}$ ($718\,\mathrm{km}$ vs $870\,\mathrm{km}$) and raises country accuracy by $5.01\,\%$ over raw GeoCLIP retrieval.

C.5 COMPUTATIONAL COMPLEXITY AND SCALABILITY BREAKDOWN

Table 13 breaks down model parameters, runtime latency, and computational share per $15\,\mathrm{s}$ clip on a single GPU. The total end-to-end processing time is $0.17\,\mathrm{s}$ per clip. By keeping foundation backbones frozen, the trainable TAP head adds only $\sim 5.1\,\mathrm{M}$ parameters (1.2% parameter overhead) and $0.015\,\mathrm{s}$ latency (9% runtime share).

Table 12: Supplementary geolocation results on GeoCLIP 1M Fibonacci gallery ($K = 5$, test $n = 2054$).

Method	Accuracy (%)				Median Dist. (km)	Branch Acc. (%)
	City	Region	Country	Continent		
GeoCLIP (Visual Prior)	4.43	19.13	46.30	77.22	870	23.0
TAP (compatibility model)	2.97	20.89	52.82	80.57	682	29.7
TAP (rerank)	4.38	21.67	51.31	80.28	718	30.3
Late Fusion (z-score calibrated)	4.48	21.67	51.85	79.94	711	29.0
GeoTAP (learned gate)	0.58	21.57	51.27	80.28	725	30.7
GeoCLAP (Audio Prior, top-1)	0.00	0.00	0.49	4.87	9184	n/a
GeoCLAP (Audio Prior, top-5)	0.00	0.00	0.44	4.67	8974	n/a
TaxaBind (Audio Prior, top-1)	0.00	0.19	2.97	11.78	8924	n/a
TaxaBind (Audio Prior, top-5)	0.00	0.00	1.07	10.61	8640	n/a

Table 13: Parameter and latency breakdown (per 15 s clip, single GPU).

Stage	Fitted?	Parameters	Latency	Share
Visual Encoder	No	304M	0.10 s	59%
Gallery Retrieval	No	—	0.02 s	12%
Audio Encoder	No	126M	0.03 s	18%
TAP Head	Yes	5.1M	0.015 s	9%
Posterior Selection	—	—	0.005 s	3%
Total	—	**435.1M**	**0.17 s**	**100%**

C.6 Disagreement analysis

Among clips where the visual prior and GeoTAP select different top candidate clusters ($n = 492$), acoustic prototype disambiguation successfully corrects visual errors in $41.8\,\%$ of instances, compared to only $14.2\,\%$ of instances where audio reranking degrades a correct visual prediction. This confirms that acoustic evidence operates as a net-positive tie-breaker when visual prior candidates are ambiguous.

C.7 Formal Formulations of Baseline Aggregation and Fusion Methods

To ensure complete reproducibility and self-containment, we provide formal mathematical formulations for all baseline frame aggregation and audiovisual fusion methods benchmarked in Table 4.

Frame Aggregation Baselines (Vision-only). Given a video clip with T sampled frames, let $\{\mathbf{m}_k\}_{k=1}^{T} \subset S^2$ denote the 3D unit direction vectors predicted by the visual backbone (GLOBE) for each frame v_k:

1. **Single frame (GLOBE mid):** Selects the prediction from the single center frame: $\hat{\mathbf{y}} = \mathbf{m}_{\lfloor T/2 \rfloor}$.
2. **Coordinate mean:** Computes the arithmetic vector mean in 3D Euclidean space $\mathbb{R}^3$ and projects to the unit sphere S^2:
$$\bar{\mathbf{m}} = \frac{1}{T}\sum_{k=1}^{T} \mathbf{m}_k, \qquad \hat{\mathbf{y}} = \frac{\bar{\mathbf{m}}}{\|\bar{\mathbf{m}}\|_2}. \tag{11}$$
3. **Spherical mean:** Computes the Fréchet directional mean on S^2:
$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in S^2} \sum_{k=1}^{T} \mathbf{y}^\top \mathbf{m}_k = \frac{\sum_{k=1}^{T} \mathbf{m}_k}{\|\sum_{k=1}^{T} \mathbf{m}_k\|_2}. \tag{12}$$
4. **Coordinate median:** Computes the component-wise scalar median across the Cartesian coordinates $\{m_{k,x}, m_{k,y}, m_{k,z}\}_{k=1}^{T}$ in $\mathbb{R}^3$:
$$\tilde{m}_c = \text{median}\left(\{m_{k,c}\}_{k=1}^{T}\right) \quad \text{for } c \in \{x, y, z\}, \qquad \hat{\mathbf{y}} = \frac{\tilde{\mathbf{m}}}{\|\tilde{\mathbf{m}}\|_2}. \tag{13}$$

5. **Spherical median (Weiszfeld):** Solves the spatial geometric median on S^2 minimizing chordal distance:

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y} \in S^2} \sum_{k=1}^{T} \|\mathbf{y} - \mathbf{m}_k\|_2. \tag{14}$$

Audiovisual Fusion Baselines. Let $\mathbf{m}_v \in S^2$ denote the visual center prediction (GLOBE mid frame), and let $\mathbf{m}_a \in S^2$ denote the independent audio retrieval prediction obtained from the Geo-CLAP audio backbone over the global gallery:

1. **Coordinate mean:** Computes equal 3D vector averaging of visual and audio predictions:

$$\hat{\mathbf{y}} = \frac{\mathbf{m}_v + \mathbf{m}_a}{\|\mathbf{m}_v + \mathbf{m}_a\|_2}. \tag{15}$$

2. **α-Weighted interpolation (α=0.99):** Performs a convex linear combination on the 3D unit vectors heavily weighted toward vision:

$$\hat{\mathbf{y}} = \frac{\alpha \mathbf{m}_v + (1-\alpha)\mathbf{m}_a}{\|\alpha \mathbf{m}_v + (1-\alpha)\mathbf{m}_a\|_2}, \qquad \text{with } \alpha = 0.99. \tag{16}$$

3. **Late score fusion ($s_v + s_a$):** For each gallery location i, computes visual logit $s_{v,i} = \mathbf{e}_v^\top \mathbf{g}_i$ and audio logit $s_{a,i} = \mathbf{a}^\top \mathbf{g}_i^{(a)}$. The fused prediction selects the maximum summed logit:

$$\hat{\imath} = \arg\max_i \left(s_{v,i} + s_{a,i}\right), \qquad \hat{\mathbf{y}} = \mathbf{m}_{\hat{\imath}}. \tag{17}$$

4. **Gated fusion (top-25% audio):** Uses audio prediction $\mathbf{m}_a$ only if the maximum audio retrieval logit $s_{a,\max}$ exceeds the 75th percentile empirical confidence threshold $\theta_{0.75}$; otherwise defaults to the visual prior $\mathbf{m}_v$:

$$\hat{\mathbf{y}} = \begin{cases} \mathbf{m}_a & \text{if } s_{a,\max} \geq \theta_{0.75}, \\ \mathbf{m}_v & \text{otherwise.} \end{cases} \tag{18}$$