

TACET: Context-Appropriate Acoustic-Social Navigation for Quadrupeds

Abstract—Quadruped robots entering hospitals, care homes, and quiet offices must be context-appropriate not only in *where* they move but in *how loudly* they move: a legged robot’s locomotion noise, dominated by foot–ground impacts, is itself a social variable. Prior social navigation respects human space but treats the robot as acoustically uniform, while quiet-locomotion methods reduce noise to an operator-specified, context-blind level. We present TACET, a context-appropriate acoustic-social navigation method that infers social context from the robot’s egocentric view and decides both where it walks and how loudly, coupling a slow fine-tuned vision-language reasoner to a fast reactive controller through a single compact behavior token, (*gait*, *speed*, *social_cost*). The same token conditions both a social costmap (*where* to go) and a quiet locomotion policy (*how loudly* to move), while a structured out-of-view memory keeps recently seen people in the reasoner’s context after they leave the camera view. On a real quadruped, context-conditioned locomotion lowers locomotion noise by up to 9.3dBA at matched speed, and across our scenarios the full method keeps personal-space compliance at 100% with low acoustic intrusion (≤ 2.9 dBA), jointly improving spatial and acoustic performance in the evaluated scenarios. The project page is available at <https://anonymous.4open.science/w/tacet>.

I. INTRODUCTION

Consider a quadruped crossing a shared building lobby (Fig. 1). A geometry-only planner treats each person as an obstacle and takes the shortest collision-free path, brushing past someone absorbed in work at a table, then threading between people standing in conversation, its footfalls fully audible. The path is safe, yet it is socially inappropriate in two distinct ways: *where* the robot moves, as it crowds personal space and cuts through a conversation, and *how loudly* it moves, as its footfalls intrude on people who need quiet. Unlike wheeled platforms, a legged robot is intrinsically loud because its locomotion noise is dominated by foot–ground impacts [1], so its acoustic footprint is a social variable that must be managed online.

Prior work largely follows two threads, each addressing only half of this problem. Social navigation makes the robot respect human space, but treats it as acoustically uniform [2], [3]; quiet locomotion makes a quadruped quieter, but to an operator-specified degree that is blind to context [1]. Acoustic-aware planning relaxes this by predicting how loud the robot sounds to nearby listeners, yet it needs a prior visual scan of the environment and manages noise at the level of the path rather than the robot’s own gait [4]. To our knowledge, no prior method has demonstrated social-context reasoning coupled to a controllable acoustic footprint: the former knows where to go but not how quiet to be, while the latter can be quiet but not when to be.

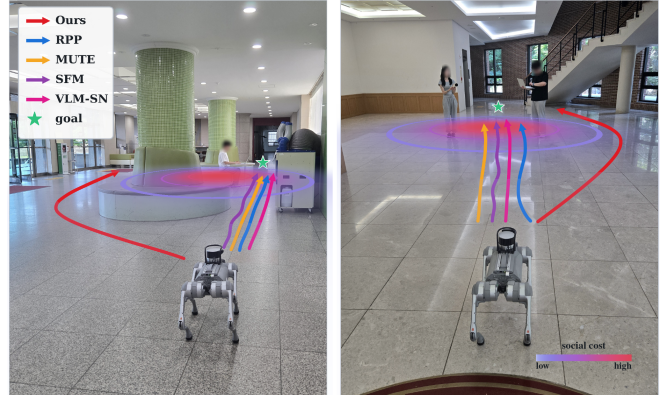


Fig. 1. Method paths on real runs of the two indoor scenarios: a focused person working at a table (left) and a standing group in conversation (right). Shading is the social costmap, rendered on the ground plane. TACET (red) swings wide and keeps its distance, whereas RPP (blue), MUTE (amber) and VLM-Social-Nav (magenta) take the short path past the people, and SFM (purple) enters the personal zone and passes closest to the seated person.

To this end, we propose **TACET** (Token-aware Acoustic-social Etiquette Traversal), a context-appropriate acoustic-social navigation method for quadruped robots that closes this gap. A slow semantic reasoner, a fine-tuned vision-language model (VLM), reads the robot’s egocentric view together with a structured memory of nearby people, including those who have just left the camera’s field of view, and emits a compact behavior token with three fields: *gait*, *speed*, and *social_cost*. A fast reactive controller consumes this token to shape both a social costmap, which governs *where* to go, and a token-conditioned quiet locomotion policy, which governs *how loudly* to move, while a standard real-time planning stack preserves safety. The two processes are coupled only through this small token interface, allowing rich but slow social reasoning to steer real-time control without entering the safety loop. Because the same token scales spatial avoidance and acoustic effort together, the robot becomes context-appropriate rather than unconditionally quiet or fast. Our main contributions are as follows:

- 1) We introduce *context-appropriate acoustic-social navigation*: a robot infers social context and modulates not only *where* but also *how loudly* it moves, unifying spatial and acoustic compliance, previously treated in isolation, under one inferred signal.
- 2) We realize it as TACET: a fine-tuned VLM emits a compact behavior token that jointly drives a social costmap with out-of-view memory and a token-conditioned quiet locomotion policy, linking slow reasoning to a real-time controller through a small, interpretable interface.

- 3) We evaluate TACET on a Unitree Go2, where it reduces locomotion noise by up to 9.3 dBA at matched speed through context-conditioned locomotion, and it is the only method to keep full Personal Space Compliance in every scenario, at a closest approach of 1.44–3.62 m against at most 1.35 m for any baseline, showing that acoustic and social compliance can be attained together rather than traded off.

II. RELATED WORK

A. Social Robot Navigation

Classical social navigation builds on geometric planners such as the social-force model [5], which are efficient and well-validated but treat people as dynamic obstacles, yielding socially blunt behavior. On the real-world SCAND dataset, geometric systems already match human demonstrations in many situations, arguing for augmenting rather than discarding them [6]. Our design inherits this view: we retain a validated real-time planning stack for reactive safety and add VLM-based semantic reasoning above it.

Learning-based methods instead imitate human navigation or optimize social objectives with reinforcement learning; Falcon, for example, predicts human trajectories and penalizes actions that obstruct them [3]. These methods motivate the proxemics-grounded metrics we adopt, but they operate kinematically in simulation and reason only about *where* to go, never the gait, foot–ground contact, or acoustics a physical quadruped must manage.

More recently, VLMs have brought open-vocabulary social reasoning to navigation. VLM-Social-Nav queries a VLM-based scoring module for a social cost term that corrects a low-level planner, structuring the VLM output into linguistic heading and speed terms mapped to linear and angular velocity [2]. Narrate2Nav instead embeds language reasoning implicitly in a visual encoder via a Barlow-Twins objective, enabling real-time RGB-only navigation that reasons in latent rather than token space [7]. In contrast to these spatial- and trajectory-level methods, which keep people’s space but assume an acoustically uniform robot and largely target wheeled platforms, TACET conditions both a social costmap and a quadruped’s quiet gait on VLM-inferred context, making the robot’s acoustic footprint itself a function of social context.

B. Quiet and Acoustic-Aware Locomotion for Legged Robots

The dominant source of quadruped locomotion noise is the foot–ground impact, whose acoustic amplitude grows with foot velocity at contact [1]. MUTE reduces it through control alone, with a quiet factor that trades noise reduction against agility and an evaluation protocol based on Mean and Peak Noise Levels (MNL, PNL) at matched actual speed [1]. The desired quietness there is an operator-specified parameter: the robot does not perceive who is nearby, and so cannot determine *when* quietness is socially required.

On the acoustic-aware navigation side, ANAVI is closest to our motivation: it visually estimates how loud the robot will sound at a listener’s location and uses this as a cost to plan

a quieter route [4]. It nonetheless depends on a prior visual scan of the environment and manages noise at the level of the path rather than the gait, and the social weighting it proposes is left unevaluated. In contrast, the token-conditioned policy in TACET is driven by VLM-inferred social context and decides online, without any prior scan of the environment, both when and how much to attenuate its own locomotion noise.

C. Lightweight VLM Reasoning for Navigation

Because large VLMs are computationally heavy and slow, a growing body of work distills their social reasoning into deployable, lightweight models. Social-LLaVA fine-tunes a LLaVA-family VLM on a human-annotated social-reasoning VQA dataset [8]; OLiVia-Nav distills context captions into a lightweight model whose embeddings select socially compliant trajectories [9]. These efforts build on egocentric and VQA resources such as SCAND, MuSoHu, JRDB-Social, and SocialNav-SUB [10]–[13]. Our reasoner shares the goal of making a large model’s social reasoning cheap enough to run in the loop, but differs in its output interface: where these produce language, continuous embeddings, or latent reasoning, we fine-tune a Qwen2.5-VL model to emit a discrete control- and costmap-conditioning token, an explicit bridge from social reasoning to both low-level locomotion and the costmap. That token conditions a social costmap and a controllable acoustic footprint on the same quadruped, without any prior scan of the environment.

III. METHODOLOGY

A. Overview

We consider a quadruped operating in a human-shared indoor space, tasked with reaching a goal pose while behaving in a context-appropriate manner: respecting the spatial social norms of nearby people (personal space, conversations, and groups) and limiting the acoustic intrusion its locomotion imposes on them. These two objectives are coupled, since how the robot should move depends on who is nearby and what they are doing, yet they act on different quantities, one spatial and one acoustic. Rather than collapsing them into a single closed-form cost, TACET infers a compact behavior token from the scene and uses it to drive a spatial mechanism and an acoustic mechanism in parallel.

TACET is organized as two processes coupled only through this token (Fig. 2). The semantic reasoner emits $\tau = \langle \text{gait}, \text{speed}, \text{social_cost} \rangle$, with $\text{gait} \in \{\text{quiet}, \text{normal}, \text{urgent}\}$, $\text{speed} \in \{\text{low}, \text{normal}, \text{high}\}$, and social_cost a scalar field amplitude at a resolution of 0.5; the reactive controller consumes it. Neither crosses into the other: the reasoner never outputs motor commands, and the controller never performs semantic reasoning.

The two processes run at different rates so slow reasoning never enters the safety loop. The reasoner emits fresh tokens at roughly 0.25 Hz and caches them, while the reactive controller—the social costmap, a global path planner, a pure-pursuit velocity controller, and the 50 Hz locomotion

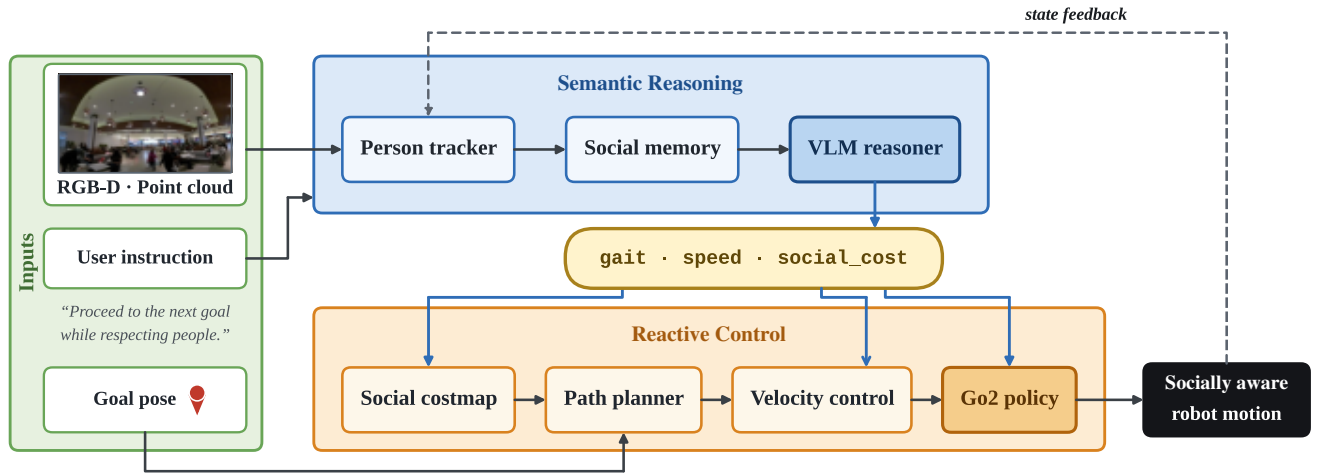


Fig. 2. Overview of TACET. The semantic reasoner (person tracker → social memory → VLM reasoner) emits the behavior token $\langle \text{gait}, \text{speed}, \text{social_cost} \rangle$; the reactive controller (social costmap → path planner → velocity controller → locomotion policy) consumes it to produce socially aware, acoustically controlled motion.

policy—runs continuously. Onboard sensing provides RGB-D, point clouds, and LiDAR-inertial odometry.

Each control cycle (Fig. 2) tracks people into a social memory that retains those out of view (Sec. III-C); the memory, the egocentric image, and the instruction go to the VLM (Sec. III-B), whose token sets the costmap (Sec. III-D) and the gait (Sec. III-E). Because the controller always acts on the most recent cached token, a delayed or dropped query degrades the robot to its previous social context rather than an unconditioned one, and the hard obstacle layer is unaffected either way.

B. Semantic Reasoner

The semantic reasoner maps an egocentric RGB image, a natural-language instruction, and the social memory to a short chain-of-thought followed by the behavior token τ . The memory is a structured record of nearby people, including those currently out of view (Sec. III-C). Producing τ as two categorical fields and one scalar amplitude makes the reasoner’s output directly parseable and consumable by the controller, in contrast to free-form language or continuous embeddings.

Auto-labeling and training data. Supervision comes from automatically generated labels. A large VLM is prompted to produce structured $\langle \text{thought}, \text{action token} \rangle$ labels on 3,950 egocentric frames drawn from MuSoHu [11], JRDB-Social [12], SCAND [10], and InCrowd-VI [14]; to emphasize these robot- and human-perspective social scenes, the auto-labeled records are repeated four times during training. They are combined with 18,951 further records: 18,042 general images from COCO [15] auto-labeled into the same record format, and 909 SocialNav-SUB frames [13] whose existing labels were mapped onto the token format, giving 34,751 training records drawn from 22,901 distinct source frames. Auto-labeling thus converts heterogeneous scenes into a single, consistent token format. Because the labeling VLM can hallucinate, every generated label is reviewed by a human annotator, who corrects samples whose reasoning or emitted token is inconsistent with the scene, so only verified

labels enter training.

Supervised fine-tuning. We fine-tune a lightweight open VLM on this mixture with parameter-efficient supervised fine-tuning: the vision encoder and language backbone are frozen and only low-rank adapters are trained, and the loss is applied to the assistant output tokens only. Training on a mixture that up-weights the egocentric sources preserves the base model’s general perception while learning the structured navigation interface. We evaluate on a held-out benchmark of 100 SCAND frames taken from sequences excluded from training and labeled by a human annotator, so the reference is independent of the VLM used for labeling. The same VLM without fine-tuning, given the same prompt, matches the reference on 30% (gait), 26% (speed), and 14% (discretized social_cost), with all three fields agreeing on 14%. After fine-tuning these rise to 95%, 96%, and 69%, with all three fields agreeing on 65% of frames. Exact agreement understates the quality of the discretized social_cost: its mean absolute error decreases from 0.879 to 0.185, and the within-one-bin (≤ 0.5) agreement increases from 47% to 96%.

Inference. A query to the reasoner takes 2.6 s at the median (95th percentile 3.5 s) on the desktop GPU of Sec. IV-A, so tokens are refreshed at roughly 0.25 Hz and cached between queries, and the high-rate controller is never blocked by VLM latency. Unlike methods that fold a VLM’s preferred action into a single planner cost term [2] or distill reasoning into continuous embeddings [9], our reasoner emits a discrete, interpretable token consumed directly by both the costmap and the policy.

C. Out-of-View Social Memory

A narrow camera field of view (FOV) means that people the robot was just reacting to often leave the frame while remaining well inside its local planning neighborhood. If the reasoner saw only the current image, an empty frame would be indistinguishable from “no one nearby,” and the behavior would collapse to a neutral gait even though a person is still close. We address this with a structured, per-person social

memory that persists nearby people across FOV loss and supplies them to the reasoner as context.

Tracking and retention. People are detected with a pose-aware human detector and localized by fusing depth with the 3D point cloud; the resulting tracks are published to the tracker-reasoner bridge, which maintains them in the odometry frame under a hold/coast/refresh policy. A confirmed track is retained while it stays within a ± 5.5 m window around the robot and its observation is recent. While observations continue, including point-cloud-only updates after the detection box disappears, position, velocity, and speed are refreshed from measurement, while orientation is updated conservatively to avoid per-frame yaw noise. Once observations stop, a moving person is coasted forward under a constant-velocity assumption and a stationary person is held in place until a time-to-live expires (12 and 14 s respectively) or the person leaves the window. A new detection within about 1.6 m of a coasting track is re-associated to it rather than spawned as a new person, so identity survives brief occlusions.

Memory record. Each retained, currently non-visible person is summarized as a structured record: a stable identity, distance and bearing relative to the robot, the time since its last observation, motion state and speed, a social context label (e.g., conversing, grouped, passing), the per-person social cost frozen at the last visible observation, the observation source, and a confidence value. Confidence starts from source quality (0.88 visual, 0.76 point-cloud-only, 0.62 coasted) and decays exponentially with observation age (time constant ≈ 18 s), which a new live observation resets. The memory therefore carries not just where a person is but how current that estimate is.

Injection into the reasoner. The active records are serialized into the VLM prompt alongside the few most recent reasoning outputs, with an explicit instruction to use the current image for visible people and to consult the memory only to decide whether an empty-looking scene is truly clear or merely occluded by the limited FOV. The memory is provided as reasoning context, never as a rule-based override: records for visible people are withheld so an empty image cannot rewrite an occluded person’s frozen cost, and the token is always emitted by the reasoner. Compared with a purely geometric buffer feeding a reactive collision-avoidance layer, our record is socially structured, consumed by the reasoner, and tags each track with an age-decayed confidence that distinguishes a recent measurement from an aging prediction.

D. Social Costmap

The social layer represents each person as an anisotropic Gaussian field aligned to the person’s heading, with a larger extent to the front for people in motion, written into one layer of a layered costmap and scaled by an amplitude and a variance. We use the standard proxemic-costmap formulation for this field [16], [17]; what TACET changes is where those two parameters come from. Prior work sets them offline, tying them to a hand-specified cue such as the person’s walking

speed [16] or a discrete activity label [18]. In TACET they follow the emitted token, so the same decision that sets how quietly the robot walks also sets how much room it leaves. Collisions remain in a separate hard obstacle layer, so social cost modulates preference without compromising geometric safety.

Per-person field. Each person induces an anisotropic 2-D Gaussian in their own frame,

$$c(\mathbf{p}) = A \exp\left(-\frac{1}{2}\left((u + \delta)^2/\sigma_{\text{long}}^2 + v^2/\sigma_{\text{side}}^2\right)\right), \quad (1)$$

with $\sigma_{\text{long}} = \sigma_{\text{front}}$ for $u \geq 0$ and σ_{rear} for $u < 0$, where u and v are the longitudinal and lateral offsets in the person frame and δ shifts the field center slightly behind the person. Splitting σ_{long} into a larger front and smaller rear extent yields the familiar egg-shaped personal space [16] oriented along the person’s travel direction for a person in motion, and a circular field otherwise.

Token $\rightarrow$ amplitude. The `social_cost` token sets the field strength. Writing s for the token value after a fixed gain and exponential-moving-average filtering, we map

$$\begin{aligned} \rho &= \max(\rho_{\text{floor}}, \min(1, s/s_{\text{max}})), \\ A &= A_0 + (A_{\text{max}} - A_0)\rho, \end{aligned} \quad (2)$$

so that more cautious situations (larger s) raise the amplitude toward A_{max} , while a floor ρ_{floor} guarantees a minimum clearance even when s is small.

Token $\rightarrow$ width and motion. The same normalized ρ , together with gait and speed multipliers M_{gait} , M_{speed} , and a per-person scale k_{shape} , sets the field width, and a motion signal $m \in [0, 1]$ elongates the front lobe for people in motion, above a small threshold m_0 :

$$\begin{aligned} \sigma &= \sigma_0(1 + 0.58\rho)M_{\text{gait}}M_{\text{speed}}k_{\text{shape}}, \\ \sigma_{\text{side}} &= \sigma_{\text{rear}} = \sigma, \end{aligned} \quad (3)$$

$$\sigma_{\text{front}} = \sigma(1 + 0.45m) \text{ if } m > m_0, \text{ else } \sigma.$$

Using the observed motion direction rather than an estimated body yaw avoids the brittleness of per-frame pose estimation while still extending cost ahead of moving people. The coefficients in (3) were tuned once in simulation and never re-tuned per scenario.

Group field. Following group-aware proxemic models, which represent an interacting group as one zone rather than a sum of individual ones [19], we merge people judged to be interacting (pairwise spacing within a small range) into a single field spanning the segment connecting them, protecting the shared interaction space (o-space):

$$c_{\text{grp}} = A_{\text{grp}} \exp\left(-\frac{1}{2}\left(d_{\perp}^2/\sigma_{\text{lat}}^2 + d_{\parallel}^2/\sigma_{\text{end}}^2\right)\right), \quad (4)$$

with $A_{\text{grp}} = \min(A_{\text{max}}, 1.25A)$, where $d_{\perp}$ is the perpendicular offset from the connecting segment, $d_{\parallel}$ the distance beyond either endpoint (zero between them), σ_{lat} a fixed lateral extent (0.65 m), and $\sigma_{\text{end}} = \max(0.35 \text{ m}, 0.35d)$ set by the inter-person spacing d . The amplitude again follows the inferred token, so the same pair induces different avoidance depending on the social context.

E. Quiet Locomotion

The gait and speed tokens select how the robot moves. Because quadruped locomotion noise is dominated by foot-ground impacts, whose acoustic energy grows with foot velocity at contact [1], we reduce perceived noise through control rather than hardware by shaping the policy’s foot-contact behavior under token conditioning.

Policy. A single policy maps a 64-dimensional observation (proprioception, the velocity command, and the token descriptor with its gait clock, excluding base linear velocity) through an MLP with hidden sizes [512, 256, 128] to twelve joint position targets at 50 Hz, tracked by joint PD control. It is trained with reinforcement learning in massively parallel simulation and deployed zero-shot on the robot. The gait and speed tokens enter the policy as physical quantities rather than a one-hot code: the gait token selects a contact schedule (a four-beat crawl for quiet; a trot with decreasing duty factor for normal and urgent) and the speed token its period, so one network covers every gait-speed pair the reasoner emits (quiet at low or normal speed, normal at any speed, urgent at high speed).

Reward. The reward combines velocity tracking with token-dependent style and acoustic terms and standard stability/smoothness regularizers. The key acoustic terms are a token-gated foot-impact penalty that tightens as the gait token moves from urgent to quiet at matched speed, together with contact-force and loading-rate penalties that cap the loudest footfalls; gait-style terms (contact-schedule tracking, clearance, and air time) keep each commanded gait well-formed; and a left-right symmetry objective, an LR joint-mirror penalty together with a yaw-rate penalty on straight-line commands, suppresses lateral drift. Unlike an operator-set quiet factor [1], the gait is chosen by a socially inferred token, so quietness is allocated by context rather than preset.

IV. EXPERIMENTS

We evaluate TACET at two levels. We first characterize the locomotion policy as a noise source in isolation, on a controlled traverse with no people present, separating the gait’s acoustic contribution from that of any detour. We then evaluate the full method in social encounters, measuring whether the robot is simultaneously space-respecting and quiet toward the people it meets. Each condition is repeated for ten trials and we report the mean, following the evaluation protocol of prior quiet-locomotion work [1].

A. Implementation Details

The platform is a Unitree Go2 with an Intel RealSense D435i RGB-D camera and a Hesai XT-16 LiDAR; VLM inference runs on a single NVIDIA RTX A6000 (48 GB). YOLOv8n-pose [20], Point-LIO [21], and Nav2’s Smac Hybrid-A* [22] run at default settings. Labels are generated with Gemini 2.5 Flash [23]; the reasoner is Qwen2.5-VL-7B [24] fine-tuned with 4-bit QLoRA [25] for one epoch. The policy is trained with PPO in Isaac Lab [26] ($\approx 25,000$ iterations, 4,096 parallel environments) and deployed through rl_sar [27]. Further implementation details, costmap and

TABLE I: Locomotion noise at matched speeds.

Speed (m/s)	Method	MNL (dBA) ↓	PNL (dBA) ↓
0.5	Default Go2	68.9 ± 2.8	72.5 ± 0.7
	MUTE ($\beta=1$)	67.8 ± 2.0	71.5 ± 0.5
	Ours	64.1 ± 1.3	65.9 ± 0.8
0.8	Default Go2	74.4 ± 2.4	77.2 ± 1.2
	MUTE ($\beta=1$)	71.7 ± 2.9	74.6 ± 0.7
	Ours	65.1 ± 2.1	68.1 ± 1.6
1.1	Default Go2	74.8 ± 1.9	76.9 ± 1.2
	MUTE ($\beta=1$)	72.3 ± 2.5	76.6 ± 0.2
	Ours	70.6 ± 2.4	75.1 ± 0.6

Speed is the commanded speed in m/s; MNL and PNL are in dBA (lower is better). Values are mean $\pm$ SD over ten trials; bold indicates best.

reward parameters, and the human-study questionnaire and protocol are on the project page.

B. Acoustic Locomotion Benchmark

Setup. Following the quiet-locomotion evaluation protocol [1], the robot traverses at three fixed commanded speeds (0.5, 0.8, and 1.1 m/s) over a path of at least 10 m, with ten trials per speed; we report per-speed results. Because here we characterize the policy as a noise *source* rather than a listener’s exposure, a calibrated sound-level meter is mounted on top of the robot at a fixed offset (0.05 m above the body) with no people present. Baselines are compared at *matched actual* speed, taken from the LiDAR-inertial odometry [21] rather than from leg odometry, whose contact assumption makes its drift gait-dependent; a method therefore cannot appear quieter merely by moving more slowly. Standing still with only its cooling fan running, the Go2 already registers about 60 dBA at the same microphone position, the floor for the locomotion levels in Table I.

Why A-weighting. A-weighting attenuates the low frequencies to which hearing is least sensitive, approximating *perceived* loudness rather than total radiated energy [28], as is conventional for quiet legged locomotion [1]. We report the Mean and Peak Noise Level (MNL, PNL), the time-averaged and maximum level over a traverse.

Baselines. We compare the default Go2 controller, our reimplementation of MUTE [1] at $\beta=1$, its quietest setting, and ours, which emits the gait token appropriate to each commanded speed (Table I).

Across all three speeds, Ours is the quietest at matched speed, lowering MNL by up to 9.3 dBA over the default controller and 6.6 dBA over MUTE (at 0.8 m/s), with a similar ordering in PNL. As expected, absolute noise rises with speed for every method, since faster foot-ground impacts are louder, but Ours stays below both baselines at each speed, showing that the per-speed gait token yields a consistent quieting margin rather than a single fixed setting. The margin is smallest at 1.1 m/s, where impact noise dominates and little headroom remains for quieting.

C. System-Level Social Navigation

Metrics. We report four objective metrics spanning the spatial-social and acoustic-social dimensions:

TABLE II: Social-navigation results across four scenarios.

Scenario	Method	PSC $\uparrow$	MinDist $\uparrow$	ΔL $\downarrow$	AID $\downarrow$
Focused (working)	RPP	95.6	1.03 ± 0.12	8.4 ± 0.5	66.3
	SFM	92.7	0.41 ± 0.05	5.5 ± 0.3	51.8
	MUTE	97.4	0.91 ± 0.14	6.4 ± 0.3	61.2
	VLM-SN	100.0	1.35 ± 0.06	5.9 ± 0.2	58.3
	Ours	100.0	3.61 ± 0.17	1.6 ± 0.2	0.0
Focused (headph.)	RPP	95.8	1.06 ± 0.14	8.1 ± 0.6	65.9
	SFM	92.2	0.39 ± 0.06	5.7 ± 0.2	52.3
	MUTE	97.1	0.89 ± 0.15	6.3 ± 0.3	60.1
	VLM-SN	100.0	1.35 ± 0.04	5.7 ± 0.2	57.9
	Ours	100.0	3.62 ± 0.19	2.9 ± 0.2	26.5
Standing group	RPP	92.3	0.90 ± 0.19	12.1 ± 1.1	100.0
	SFM	92.8	1.02 ± 0.02	7.8 ± 0.6	73.3
	MUTE	93.7	1.13 ± 0.05	4.9 ± 0.5	42.9
	VLM-SN	90.4	0.87 ± 0.16	7.3 ± 0.5	65.7
	Ours	100.0	2.25 ± 0.21	2.7 ± 0.3	22.3
Moving person	RPP	87.7	0.59 ± 0.05	7.9 ± 0.7	83.2
	SFM	90.2	0.75 ± 0.04	5.1 ± 0.5	51.3
	MUTE	98.9	1.21 ± 0.11	3.8 ± 0.3	24.7
	VLM-SN	94.1	0.69 ± 0.07	4.1 ± 0.4	29.4
	Ours	100.0	1.44 ± 0.12	1.7 ± 0.3	0.0

PSC in %, MinDist in m, ΔL (sound-level increment) in dBA, AID in %; VLM-SN: VLM-Social-Nav [2]. Means over ten trials; MinDist and ΔL are mean $\pm$ SD. Bold indicates best.

- **Personal Space Compliance (PSC, %, $\uparrow$).** Fraction of the run for which the robot keeps a center-to-center distance of at least 1.2m from every person, which is Hall’s *personal/social* boundary [29] and a standard proxemic proxy [3], [30].
- **Minimum Distance (MinDist, m, $\uparrow$).** The smallest center-to-center robot–person distance reached during the run. PSC measures how *long* the robot stays outside the personal zone, MinDist how *deeply* it intrudes at the closest point; a run can score well on one and poorly on the other, so we report both [30].
- **Sound-Level Increment (ΔL , dBA, $\downarrow$).** With L_{on} the A-weighted level at the listener while the robot operates and L_{amb} the matched ambient level, we report $\Delta L = L_{\text{on}} - L_{\text{amb}}$. It is the increase in total listener level due to robot operation, not an estimate of robot-only SPL. The meter sits at the person’s location; L_{amb} is 49 dBA in the two focused-person conditions, 47 dBA for the group, and 53 dBA outdoors.
- **Acoustic Intrusion Duration (AID, %, $\downarrow$).** Fraction of the run for which ΔL exceeds 5 dBA, the level at which a change is clearly noticeable rather than barely perceptible [31], so AID captures how long the robot is *audibly* above ambient.

Scenarios. Seven encounters test context-dependent behavior. The first four place people on the route and use the social metrics above; the last three do not.

- **Focused person, working (indoor).** A stationary worker in a quiet indoor space tests whether the robot avoids acoustic intrusion.
- **Focused person, headphones (indoor).** The same scene with only headphones added isolates a cue that relaxes acoustic, but not spatial, caution.
- **Standing group (indoor).** A conversing group blocks

TABLE III: Time to goal and emitted tokens by scenario.

Scenario	Time to goal (s)			Emitted token		
	mean	best	Ours	gait	spd	cost
Focused (working)	31.8	21.2	36.8	quiet	low	4.0
Focused (headph.)	31.8	22.0	33.8	normal	low	4.0
Standing group	17.2	13.6	25.6	quiet	low	3.5
Moving person	16.7	14.8	21.8	quiet	normal	2.5
Empty corridor	16.2	14.4	14.4	normal	normal	1.5
Sports field	18.3	14.4	14.0	normal	high	1.0
Emergency	15.4	14.2	11.2	urgent	high	1.0

Times are trial means; tokens are trial modes. “Baseline mean” averages RPP, SFM, MUTE, and VLM-Social-Nav, whereas “Baseline best” is the fastest of those four; Ours is excluded from both baseline summaries. Bold marks the fastest entry in each row. The midrule separates person-facing and no-person scenarios.

the route, requiring the robot to avoid its shared space.

- **Moving person (outdoor).** A crossing pedestrian tests yielding under higher ambient noise.
- **Empty corridor (indoor).** With no person present, any detour or slowdown is overhead.
- **Open sports field (outdoor).** Distant activity leaves the robot’s route unobstructed.
- **Emergency (indoor).** A flashing red siren in an empty, dim corridor tests whether urgency overrides acoustic restraint.

Baselines. All baselines run on the same Go2 and planning stack. **RPP:** the Nav2 Regulated Pure Pursuit controller treats people as ordinary obstacles with the default gait. **SFM:** a classical social planner whose proxemic repulsion shapes the path [5], default gait. **MUTE:** the quiet locomotion policy driven by the RPP controller, with no social reasoning [1]. **VLM-Social-Nav:** a VLM-based social planner that scores candidate actions with a vision-language model [2], using GPT-4o as the reasoning model, default gait; it reasons about social context but not about the robot’s own acoustic footprint. **Ours** is the complete TACET method.

Table II reports the per-scenario results. Ours is the only method that keeps PSC at 100% in all four scenarios; VLM-Social-Nav reaches 100% with the focused person but not with the group or the moving person. Ours also keeps a far wider margin at the closest approach, with MinDist never below 1.44 m against at most 1.35 m for any baseline.

All runs reached the goal without contacting a person, but not always because the robot made room. SFM oscillated in the narrow passage around the focused person, scraping the wall and nearly brushing the person; RPP did not yield to the moving person and stayed clear only because the person stepped aside. Three baseline conditions came within arm’s reach of a person: SFM in both focused-person conditions (MinDist 0.41 and 0.39 m) and RPP with the moving person (0.59 m). Ours is also the quietest in every scenario and audible above ambient for the smallest fraction of the run.

What the reasoner adds. The two focused-person conditions hold the scene fixed and change only the headphones. Methods without token-conditioned gait control behave near-identically in both; VLM-Social-Nav, for example, yields ΔL 5.9 vs. 5.7 dBA. Ours separates them while maintaining essentially the same clearance (MinDist 3.61 vs. 3.62 m)

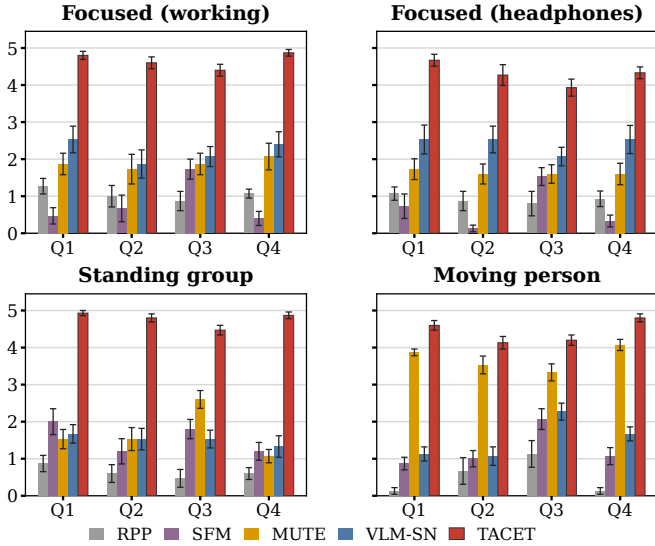


Fig. 3. Per-question results of the blinded video study. Q1: safe and comfortable distance; Q2: contextual consideration; Q3: footstep-noise appropriateness; Q4: scenario-specific etiquette. Bars show mean scores on a 0–5 Likert scale; error bars show SEs. Higher is better.

and relaxing the acoustic effort where the listener is less sensitive to it (ΔL 1.6 vs. 2.9 dBA, AID 0.0 vs. 26.5%). With the closest approach unchanged and only the gait token differing, the acoustic difference is consistent with the gait effect measured in Table I (Sec. IV-E).

Table III shows that the cost of this behavior varies with what the context calls for: where people need space or quiet, TACET spends at most 1.5 times the mean baseline time, but in an empty corridor it matches the fastest baseline (14.4 s) and on an open sports field it is the fastest method (14.0 s against 14.4 s), as it is with the siren light in the corridor (11.2 s against 14.2 s). The token columns show why: the instruction names neither gait, speed, nor clearance, so the token follows the scene and the behavior follows the token.

D. Human Study

Following prior social-navigation user studies [2], [6], we ran a blinded video study with $N=15$ observers. Each observer rated the 20 clips formed by four person-facing scenarios and five methods, presented in randomized order. After each clip, observers rated Q1–Q4 on a 0–5 Likert scale.

Fig. 3 plots the per-question averages, taken over paired within-observer ratings, with standard errors for the five methods in each scenario. TACET receives the highest agreement in all sixteen scenario–question cells, averaging 4.54 against 2.23 for MUTE, 1.92 for VLM-Social-Nav, 1.07 for SFM and 0.78 for RPP, and its standard errors are the smallest (0.14 on average against 0.29 for VLM-Social-Nav), indicating more consistent ratings across observers. The margin narrows only with the moving person, where MUTE reaches 3.33–4.07; that is also where MUTE is the strongest baseline on the objective metrics (Table II), consistent with its observer ratings there.

E. Ablation Study

We remove one component of TACET at a time, keeping the rest fixed, and repeat this on two representative person-

TABLE IV: Ablations on two person-facing scenarios.

Scenario	Variant	PSC $\uparrow$	MinDist $\uparrow$	ΔL $\downarrow$	AID $\downarrow$
Focused (working)	Qwen2.5-VL-7B	100.0	1.35 ± 0.10	4.5 ± 0.5	31.8
	InternVL3.5-8B	100.0	1.33 ± 0.12	4.9 ± 0.4	35.1
	Ours (full)	100.0	3.61 ± 0.17	1.6 ± 0.2	0.0
	w/o soc. layer	96.5	1.18 ± 0.11	4.3 ± 0.3	29.3
	w/o trained pol.	100.0	3.47 ± 0.19	7.4 ± 0.5	63.3
	w/o soc. memory	100.0	3.54 ± 0.24	3.9 ± 0.3	18.7
Moving person	Qwen2.5-VL-7B	100.0	1.38 ± 0.11	3.5 ± 0.5	36.1
	InternVL3.5-8B	100.0	1.33 ± 0.12	3.8 ± 0.7	39.3
	Ours (full)	100.0	1.44 ± 0.12	1.7 ± 0.3	0.0
	w/o soc. layer	94.7	0.89 ± 0.12	3.8 ± 0.3	37.4
	w/o trained pol.	100.0	1.39 ± 0.13	4.2 ± 0.5	41.5
	w/o soc. memory	100.0	1.42 ± 0.09	2.6 ± 0.4	19.3

PSC and AID in %, MinDist in m, ΔL in dBA; means over ten trials; MinDist and ΔL are mean $\pm$ SD; bold indicates best. Rows above each rule swap only the reasoner that emits τ ; rows below remove one component.

facing scenarios (Table IV). Removing the social layer disables (1)–(4) while the trained policy is kept, so people are planned around as ordinary obstacles; removing the trained policy (default gait, with the social layer and social memory intact) is its mirror image, keeping the socially aware detour but not the token-conditioned policy. Because people pass in and out of the camera FOV as the robot moves, the *w/o social memory* variant isolates whether behavior stays consistent across these brief occlusions. Separately, the top block of each scenario keeps the architecture fixed and swaps only the reasoner that emits τ , isolating the contribution of fine-tuning beyond prompt engineering.

Removing the social layer costs both compliance and clearance in each scenario, dropping MinDist inside the 1.2 m zone the full method never enters. The acoustic terms degrade with it: with people treated as ordinary obstacles the robot passes closer, so the same gait is louder at the listener. Removing the trained policy instead leaves the route essentially unchanged—both MinDist differences fall within one standard deviation—while ΔL and AID degrade the most of any variant, supporting an acoustic contribution from the gait policy beyond that of the detour. Removing social memory alters the route only marginally (3.54 and 1.42 m), yet the robot is audible above ambient for 18.7% and 19.3% of the run against 0.0% for the full method: without the memory a person who briefly leaves the camera view is no longer carried in the reasoner’s context, so the policy is no longer conditioned on a quiet token across the occlusion.

Swapping the reasoner degrades both axes. With the focused person, a prompted Qwen2.5-VL-7B [24] keeps only 1.35 m of clearance against 3.61 m for the full method, close to what removing the social layer gives, and raises ΔL to 4.5 dBA with AID at 31.8%; InternVL3.5-8B [32] behaves the same way (1.33 m, 4.9 dBA, 35.1%). With the moving person the route changes far less (1.38 and 1.33 m against 1.44 m), yet the robot is audible above ambient for 36.1% and 39.3% of the run against 0.0% for the full method. Since the architecture is unchanged, the loss follows from the token interface itself rather than from any component, consistent with the base model’s low token agreement in Sec. III-B.

These variants also separate distance from gait. Because

Ours both leaves more room and walks more quietly, its lower ΔL in Table II might be read as a by-product of the larger MinDist. The *w/o trained policy* variant keeps the closest approach within one standard deviation of the full method yet raises ΔL by 5.8 and 2.5 dBA. The cleanest isolation is Table I, which removes geometry altogether (the meter rides on the robot and no one is present), yet gaits still differ by up to 9.3 dBA at matched speed. ΔL is integrated over the whole run rather than read at the closest approach, so a shared MinDist does not by itself fix the acoustic outcome.

V. CONCLUSION

We presented TACET, which treats a quadruped's footstep noise as something the robot decides rather than a fixed property of its hardware. One inferred token sets both where it walks and how loudly, improving both spatial and acoustic metrics in the evaluated scenarios, and the ablations support an acoustic contribution from the gait policy beyond that of the detour. Quietness is therefore allocated by context rather than preset, with the reasoner emitting a quiet gait near a focused worker and an urgent one in the emergency scenario.

Limitations include the reasoner's low update rate, which the reactive costmap partially mitigates but which we do not evaluate under rapid scene changes. Socially aware detours also increase time to goal, and we do not evaluate their effect on people sharing the space. Finally, the scenario set is limited. Future work targets faster onboard reasoning and broader evaluation in more dynamic, populated settings.

REFERENCES

- [1] Z. Cao, B. Nie, Y. Zhang, and Y. Gao, "Minimizing acoustic noise: Enhancing quiet locomotion for quadruped robots in indoor applications," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2025, pp. 17972–17979.
- [2] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha, "VLM-Social-Nav: Socially aware robot navigation through scoring using vision-language models," *IEEE Robot. Autom. Lett.*, vol. 10, no. 1, pp. 508–515, 2025.
- [3] Z. Gong, T. Hu, R. Qiu, and J. Liang, "From cognition to precognition: A future-aware framework for social navigation," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2025, pp. 9122–9129.
- [4] V. Jain, R. Veerapaneni, and Y. Bisk, "ANAVI: Audio noise awareness using visual of indoor environments for NAVigation," in *Proc. Conf. Robot Learning (CoRL)*, vol. 270, 2025, pp. 3924–3942.
- [5] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51, no. 5, pp. 4282–4286, 1995.
- [6] A. H. Raj, Z. Hu, H. Karnan, R. Chandra, A. Payandeh, L. Mao, P. Stone, J. Biswas, and X. Xiao, "Rethinking social robot navigation: Leveraging the best of two worlds," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024, pp. 16330–16337.
- [7] A. Payandeh, A. Pokhrel, D. Song, M. Zampieri, and X. Xiao, "Narrate2Nav: Real-time visual navigation with implicit language reasoning in human-centric environments," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2026.
- [8] A. Payandeh, D. Song, M. Nazeri, J. Liang, P. Mukherjee, A. H. Raj, Y. Kong, D. Manocha, and X. Xiao, "Social-LLaVA: Enhancing social robot navigation through human-language reasoning," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2025, pp. 17192–17198.
- [9] S. Narasimhan, A. H. Tan, D. Choi, and G. Nejat, "OLiViA-Nav: An online lifelong vision language approach for mobile robot social navigation," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2025, pp. 9130–9137.
- [10] H. Karnan *et al.*, "Socially compliant navigation dataset (SCAND): A large-scale dataset of demonstrations for social navigation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 11807–11814, 2022.
- [11] D. M. Nguyen, M. Nazeri, A. Payandeh, A. Datar, and X. Xiao, "Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2023, pp. 7442–7447.
- [12] S. Jahangard, Z. Cai, S. Wen, and H. Rezatofighi, "JRDB-Social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 22087–22097.
- [13] M. J. Munje, C. Tang, S. Liu, Z. Hu, Y. Zhu, J. Cui, G. Warnell, J. Biswas, and P. Stone, "SocialNav-SUB: Benchmarking VLMs for scene understanding in social robot navigation," in *Proc. Conf. Robot Learning (CoRL)*, vol. 305, 2025, pp. 1120–1143.
- [14] M. Bamdad, H.-P. Hutter, and A. Darvishy, "InCrowd-VI: A realistic visual-inertial dataset for evaluating simultaneous localization and mapping in indoor pedestrian-rich spaces for human navigation," *Sensors*, vol. 24, no. 24, p. 8164, 2024.
- [15] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2014, pp. 740–755.
- [16] R. Kirby, "Social robot navigation," Ph.D. dissertation, Robotics Institute, Carnegie Mellon University, 2010, tech. Rep. CMU-RI-TR-10-13.
- [17] D. V. Lu, D. Hershberger, and W. D. Smart, "Layered costmaps for context-sensitive navigation," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2014, pp. 709–715.
- [18] J. Ginés Clavero, F. Martín Rico, F. J. Rodríguez-Lera, J. M. Guerrero Hernández, and V. Matellán Olivera, "Defining adaptive proxemic zones for activity-aware navigation," in *Advances in Physical Agents II (WAF 2020)*, vol. 1285, 2021, pp. 3–17.
- [19] R. M. de Sousa, D. Barrios-Aranibar, J. Diaz-Amado, R. E. Patiño-Escarcina, and R. M. P. Trindade, "A new approach for including social conventions into social robots navigation by using polygonal triangulation and group asymmetric Gaussian functions," *Sensors*, vol. 22, no. 12, p. 4602, 2022.
- [20] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023, <https://github.com/ultralytics/ultralytics>.
- [21] D. He, W. Xu, N. Chen, F. Kong, C. Yuan, and F. Zhang, "Point-LIO: Robust high-bandwidth light detection and ranging inertial odometry," *Adv. Intell. Syst.*, vol. 5, no. 7, p. 2200459, 2023.
- [22] S. Macenski, F. Martín, R. White, and J. Ginés Clavero, "The Marathon 2: A navigation system," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2020, pp. 2718–2725.
- [23] Gemini Team, Google, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.
- [24] S. Bai *et al.*, "Qwen2.5-VL technical report," *arXiv preprint arXiv:2502.13923*, 2025.
- [25] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 10088–10115.
- [26] M. Mittal, P. Roth, J. Tigue, A. Richard, *et al.*, "Isaac Lab: A GPU-accelerated simulation framework for multi-modal robot learning," *arXiv preprint arXiv:2511.04831*, 2025.
- [27] Z. Fan, "rl_sar: Simulation verification and physical deployment of robot reinforcement learning algorithm," 2024, https://github.com/fan-ziqi/rl_sar.
- [28] International Electrotechnical Commission, "Electroacoustics—sound level meters—part 1: Specifications," IEC, International Standard 61672-1, 2013.
- [29] E. T. Hall, *The Hidden Dimension*. Doubleday, 1966.
- [30] A. Francis *et al.*, "Principles and guidelines for evaluating social robot navigation algorithms," *ACM Trans. Human-Robot Interaction*, vol. 14, no. 2, pp. 34:1–34:65, 2025.
- [31] U.S. Federal Highway Administration, "Highway traffic noise: Analysis and abatement guidance," FHWA, Tech. Rep. FHWA-HEP-10-025, 2011.
- [32] W. Wang *et al.*, "InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.