THERMAL PLUME ATTACK: PHYSICS-INSPIRED ATTACKS ON INFRARED VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared vision-language models (IR-VLMs) support classification, image captioning, and visual question answering in low-visibility environments, yet their robustness to non-contact thermal variations remains underexplored. We propose *Thermal Plume Attack*, a physics-inspired black-box framework that models a rising thermal plume and optimizes its configuration with CMA-ES. A shared temperature field couples refractive displacement with radiance variation, yielding a structured thermal perturbation. Experiments on person/vehicle crops from the FLIR and LLVIP datasets show effective top-1 flips for LanguageBind Thermal and IRGPT. Evaluations across four CLIP backbones adapted to infrared data and matched controls further reveal that this sensitivity is not confined to a single scorer or architecture. These results establish Thermal Plume Attack as a controlled testbed for studying IR-VLM behavior under structured thermal variations.

1 INTRODUCTION

Infrared vision-language models (IR-VLMs) are increasingly used for classification, image captioning, and visual question answering in low- visibility settings. Their visual evidence is formed from emitted thermal radiation, which makes infrared robustness distinct from conventional RGB robustness. It remains unclear whether a structured, non-contact thermal variation can systematically change IR-VLM decisions.

Three protocol issues make this question easy to misread. A non-contact thermal change must affect both image formation and semantic evidence, so an image-space score alone does not explain the physical mechanism. Generative IR-VLMs expose no differentiable objective, and a target-specific flip must be separated from an open-set or parser-derived event. Finally, clean-correct gating, scorer-specific budgets, and direction-specific denominators must remain fixed across comparisons.

We study this question with Thermal Plume Attack. The framework renders a rising plume from a shared temperature field, couples refractive displacement with radiance variation, and optimizes a compact parameterization with CMA-ES. We evaluate this structured non-contact variation directly on IR-specialized scorers using person/vehicle crops, complementing prior studies that focus on target-attached artifacts or generic scorers. AirflowAttack (Su et al., 2026) is a concurrent study with a different target model class and perturbation parameterization.

As shown in Figure 1, the framework has three stages: target-conditioned attack setup, physics-inspired plume rendering, and black-box parameter optimization. We evaluate direct attacks on LanguageBind Thermal and IRGPT, then analyze four IR-adapted CLIP encoders under a model-specific protocol. Generic CLIP is retained only as a secondary transfer diagnostic.

The contributions are:

- We propose Thermal Plume Attack, a physics-inspired plume renderer coupled with a compact, score-only CMA-ES optimization framework.
- We evaluate the framework across IR-specialized VLMs, IR-adapted CLIP encoders, and open-ended diagnostics on FLIR ADAS and LLVIP, establishing consistent attack effects across models and tasks.
- We analyze model sensitivity and optimization behavior, including transfer and holdout tests, to characterize how structured thermal variations alter infrared multimodal decisions.

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.

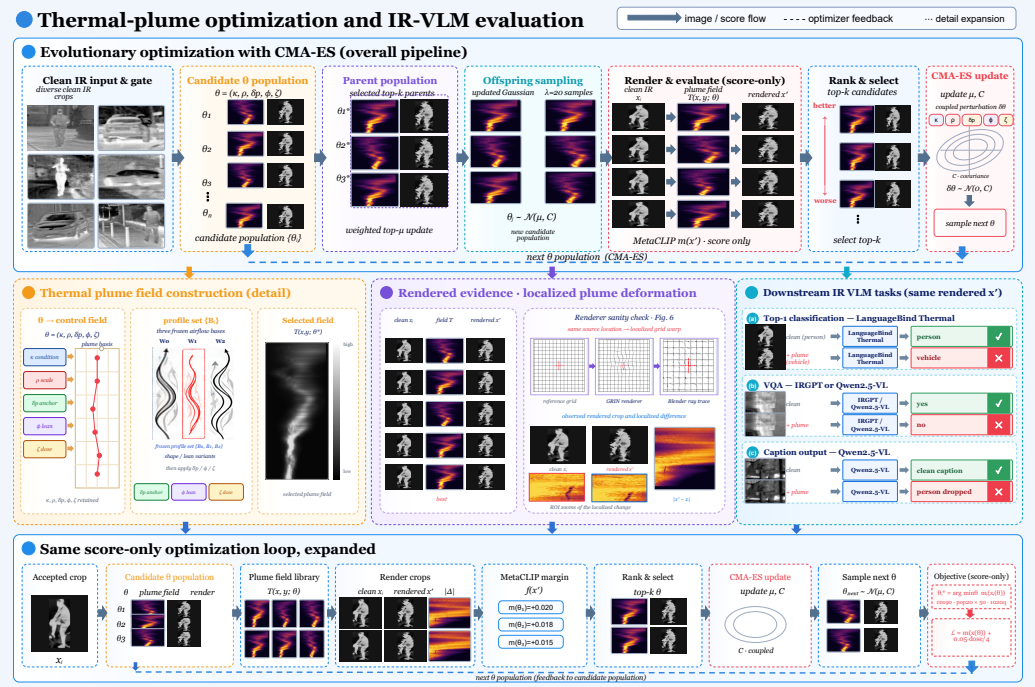

Figure 1: Thermal-plume optimization and IR-VLM evaluation. Clean and adversarial thermal crops are evaluated by LanguageBind Thermal, IRGPT, and IR-adapted CLIP scorers; the panels report target flips and target-specific attack success rates.

2 Related Work

2.1 Infrared Vision–Language Models

CLIP-style contrastive pretraining has been extended to remote-sensing and infrared imagery through RemoteCLIP (Liu et al., 2024) and GeoRSCLIP, together with its RS5M dataset (Zhang et al., 2024). Generic VLMs such as LLaVA (Liu et al., 2023), InstructBLIP (Dai et al., 2023), and Qwen2.5-VL (Bai et al., 2025) have also been applied to thermal data. LanguageBind Thermal (Zhu et al., 2024) and IRGPT (Cao et al., 2025) learn from infrared data directly, but their adversarial robustness has not been studied systematically.

2.2 Adversarial Attacks in the RGB Domain

RGB attacks span white-box optimization (FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018), C&W (Carlini and Wagner, 2017), and AutoAttack (Croce and Hein, 2020)), query-based and universal perturbations (Ilyas et al., 2018; Moosavi-Dezfooli et al., 2017), and transfer methods based on momentum or input diversity (Dong et al., 2018; Xie et al., 2019); AdvCLIP (Zhou et al., 2023) and Co-Attack (Zhang et al., 2022) target multimodal encoders directly. These attacks assume RGB reflectance images or generic CLIP features and do not constrain perturbations by thermal optics or emitted radiance.

2.3 Physical Attacks in the Infrared Domain

Contact-based methods alter the target's own heat signature (HOTCOLD (Wei et al., 2023), camouflage clothing (Zhu et al., 2022), Fourier-shaped patches (Yong et al., 2026), bimodal clothing (Long et al., 2026), stripe patterns (Zheng et al., 2026)) and require material on or near the target. A second line attacks infrared VLMs with template or statistical priors (UCGP (Hu et al., 2026), InfraQR (Li et al., 2026), QR-STT (Chen et al., 2026)); AirflowAttack (Su et al., 2026) is the closest concurrent study, using a generator-based universal airflow prior for a generic CLIP scorer, whereas we use

explicit plume dynamics with direct attacks on IR-specialized scorers. Relative to QR-STT, which studies structured thermal triggers and their transfer to captioning and VQA, our focus is a shared temperature field and score-only per-image optimization of a physically parameterized plume. Turbulence treatments such as P2S (Mao et al., 2021) and TurbuGAN (Feng et al., 2022) model atmospheric variation as an imaging degradation; we instead render the attack from an explicit synthetic plume model. Table 7 in Appendix B summarizes the differences. We use these comparisons to position the benchmark relative to infrared physical and atmospheric attack settings.

3 METHOD

3.1 FRAMEWORK OVERVIEW

Given a thermal crop x, its target box b, a frozen manifest, and a score-returning model interface, the pipeline first applies clean-correct gating, then selects a frozen plume condition, renders a coupled displacement–radiance field, optimizes the five-dimensional θ with score-only CMA-ES, and records target-specific and open-set outcomes. Figure 1 summarizes these stages. The renderer and optimizer are shared across scorers; only the scorer, prompt set, query budget, and LPIPS weight change by protocol.

3.2 THREAT MODEL

The attacker operates in a black-box setting: it may query an IR-specialized scorer (the primary target), an IR-adapted CLIP scorer, or a generic CLIP scorer used for a secondary transfer diagnostic, but has no parameters, gradients, or training data. Each image is optimized independently. Query budgets differ by scorer class because generative inference is expensive, and are stated wherever a number is reported: CLIP-class scorers (LanguageBind Thermal, the IR-adapted variants, and the generic encoders) use the frozen black-box protocol defined in the implementation details; the open-ended caption/VQA diagnostics in Appendix D evaluate plumes optimized for the LanguageBind scorer and therefore require no additional optimization queries. The scorer roles are defined in the surrounding text and reproducibility appendix.

For evaluation we use the ground-truth target box as an oracle localization signal. The box is held fixed, so the perturbation stays beside rather than on the target. Queries return a continuous margin, which is the signal optimized by CMA-ES: cosine text–image scores for the CLIP-family scorers, and teacher-forced label sequence log-probabilities for the generative IRGPT scorer. We therefore assume a score-returning interface (as in standard score-based black-box attacks) rather than text-only decoding; the attacker still has no gradients, parameters, logits beyond the required score, or training data.

The headline result for each direct target is a target-specific top-1 flip on its clean-correct subset. For IRGPT, we additionally report the broader open-set event in which the source label is no longer top-1, including outcomes assigned to "other." Gating, denominators, and the separate open-ended VQA diagnostic are defined in Section 4.1.

3.3 EXPLICIT THERMAL-PLUME RENDERING

Rather than optimizing raw pixels, we generate a perturbation field from an explicit thermal-plume model. For every image, CMA-ES optimizes the fixed five-parameter vector

$$\theta = (\kappa, \rho, \delta_p, \phi, \zeta), \tag{1}$$

where κ is the discrete plume-condition index, ρ is the dimensionless scale, δ_p is the target-relative anchor offset in pixels, ϕ is the plume lean angle in degrees, and ζ is the ROI-p95 displacement dose in pixels. We reserve α exclusively for the LPIPS weight used in the optimization objective.

The mean field follows the Morton–Taylor–Turner (MTT) integral model. We add Lane–Glindemann–Dainty FFT turbulence, Gifford meandering, and six independently perturbed filaments. Implementation constants, equations, and fidelity checks are given in Appendix A.

3.4 Refraction and Radiance Rendering

The renderer maps temperature to refractive displacement via the Gladstone–Dale relation and, more heuristically, to radiance via a temperature-informed emission model. The displacement field is integrated along the plume chord and rescaled so that its ROI-p95 equals the dose parameter. Anchors span 0.18–0.70 px for short-focal small-flow sources and 2.5–4.6 px for surveillance-scale exhaust trails; detailed measurements are in Appendix A.

Radiance rendering combines a white-hot overlay with temperature-field texture, using the same plume field that drives refraction. Implementation details and the resulting physical anchors are provided in Appendix A.

Attack mechanism. The same temperature field drives both terms of the perturbation. Across each line of sight, the field is integrated into a smooth refractive displacement and then rescaled to the requested ROI-p95 dose; the temperature residual is rendered as a coupled radiance cue. Thus the attack changes local geometry and thermal appearance through one shared plume state, while CMA-ES searches only the five configuration variables. Figure 2 shows the field, three horizontal sections, the warped grid, and the resulting image response.

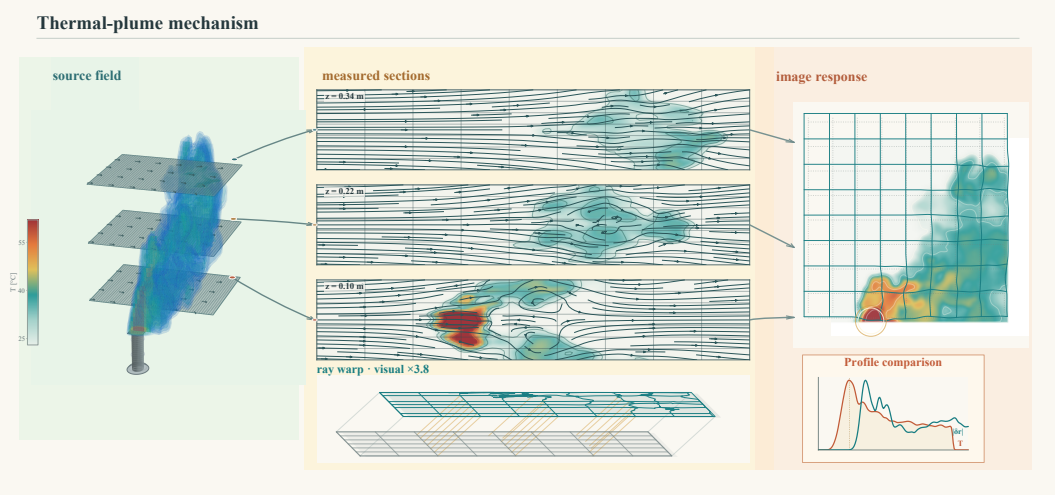

Figure 2: Thermal-plume mechanism. The source field is sampled at three horizontal sections, where the local field and flow lines produce a localized ray warp and the corresponding image response. The lower strip visualizes the warped grid, and the side profile compares the temperature and displacement traces.

3.5 Attack Optimization

The optimization domain is the bounded vector $\theta = (\kappa, \rho, \delta_p, \phi, \zeta)$ defined above. CLIP-class scorers and LanguageBind Thermal use CMA-ES with a population of 20 for 50 iterations (nominal 1,000 queries/image; 1,020 total scorer calls are recorded because the implementation adds one initial validation, identically for every method); the direct IRGPT run uses the same optimizer with $20 \times 25 = 500$ queries/image. Bounds, seeds, and the condition-index definition are listed in Appendix C. The objective is

$$\min_{\theta} m(x_\theta) + \alpha \, \mathrm{LPIPS}(x, x_\theta) + \beta \frac{\zeta}{d_{\max}}, \tag{2}$$

where m is the scorer margin of the model under attack and ζ is the ROI-p95 displacement dose in pixels. Here β is the fixed dose-penalty coefficient and $d_{\max} = 4\,\mathrm{px}$ is the normalization scale, chosen as the upper bound of the search range $\zeta \in [0.5, 4.0]$ so that the dose term is dimensionless and bounded by β. The main results use $\alpha = 0$, the naturalness trade-off uses $\alpha = 0.2$, and $\beta = 0.05$; the same budget–α pairing is used for every method in a comparison. Optimization uses score access and no gradients.

3.6 SCORERS

LanguageBind Thermal is scored by the cosine similarity difference between a frozen person prompt ("a thermal image of a pedestrian") and a frozen vehicle prompt ("thermal vehicle"), $m = s_{\text{person}} - s_{\text{vehicle}}$; its clean-correct subset is defined by $m > 0$ on the clean crop. The IR-adapted and generic CLIP encoders use the corresponding prompt sets listed in Appendix C; their margin is the mean person score minus the mean vehicle score over the prompt set.

IRGPT is an InternVL2-8B model fine-tuned on infrared data. Because its output is a generated sequence, we score it deterministically by teacher-forced label log-probabilities under the frozen prompt "What is in this thermal image? Choose exactly one: person, vehicle, other." Let $\log P(\ell)$ be the candidate-sequence score of continuing the prompt with label ℓ (raw teacher-forced log-probabilities; the candidates are compared directly, so no softmax normalization over labels is applied). For source class s and target class t, the optimized sequence margin is

$$m_{\text{IRGPT}} \;=\; \log P(s) - \log P(t), \tag{3}$$

computed by first-token and teacher-forced label log-probabilities (no sampling). A target-specific flip is recorded only when t becomes the unique top-1 label among person, vehicle, and other. We separately count open-set flips, where s is no longer top-1 but the winner may be either t or other. The positive $\beta\zeta/d_{\max}$ term discourages high-dose solutions during optimization and cannot inflate either count.

3.7 EVALUATION PROTOCOL

For the binary surrogate criterion, a flip is recorded when

$$f_{\text{clip}} = s_{\text{person}} - s_{\text{vehicle}} \leq 0. \tag{4}$$

For the expanded prompt-set top-1 criterion, we use

$$f_{\text{exp}} = s_{\text{person}} - \max(s_{\text{vehicle}}, s_{\text{other}}) \leq 0, \tag{5}$$

where s_{other} is the maximum score over five "other" prompts. Because

$$\max(s_{\text{vehicle}}, s_{\text{other}}) \geq s_{\text{vehicle}}, \tag{6}$$

every binary flip ($f_{\text{clip}} \leq 0$) is also a three-class flip ($f_{\text{exp}} \leq 0$), so the binary criterion is the more conservative headline rate and the three-class rate is reported as an upper bound in Appendix D. We report the flipped-target distribution separately for vehicle and other outcomes. The binary criterion is a surrogate and does not by itself establish vehicle-specific relabeling.

For the direct IR targets, LanguageBind uses the person and vehicle prompts frozen by the gate, and IRGPT uses the directional sequence margin above; both are optimized with the same score-only black-box procedure under the frozen protocol. Because these forced-choice scores are closed-set surrogates, we additionally probe open-ended behavior with a caption/VQA protocol in Appendix D: there a flip is a clean "yes" answer to "Is there a person in this thermal image?" becoming "no" after the attack. The two protocols have different denominators and answer spaces, and only target-specific top-1 numbers are used as headline ASRs.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Prompt details are deferred to the appendix; this paragraph defines the pool and the gate. The DEV pool contains 714 single-person windows (214 FLIR ADAS crops from 10 recordings, 500 LLVIP crops from 26 sequences) and 517 vehicle windows (ADAS). A crop is *clean-correct* (gate-pass) if the frozen prompts' top-1 label on the clean crop equals the ground-truth class (for the binary person–vehicle pair, a strictly positive CLIP margin). Only clean-correct images are attacked, so every ASR denominator equals the gate-pass count of the scorer under attack. Full-pool IRGPT evaluation. IRGPT is an 8B generative model, so its direct runs use 500 queries per image, but both directions cover the complete frozen DEV manifests. The person and vehicle gates pass 658/714 and 468/517 crops, respectively; every gate-pass crop is attacked once with the frozen prompt and

per-image CMA seed. LanguageBind's corresponding denominators are 499 person crops and 517 vehicle crops, as shown by the fractions in Table 2; no denominator is pooled across scorers. TEST-2 v2 is a frozen, independent 60-image LLVIP holdout across 26 sequences (Appendix G). Runtime is about 60 s/image for LanguageBind and 35 s/image for the CLIP encoders at 1,000 queries, and about 160 s/image for IRGPT's 500-query run. The direct IR targets and generic references use $\alpha = 0$ with frozen prompts, whereas the IR-adapted analysis uses per-model prompts and $\alpha = 0.2$; their ASRs are therefore interpreted within role rather than pooled into one robustness leaderboard.

Reproducibility-critical settings. All scorers use the same 40 px-padded crop, bilinear 512×640 8-bit conversion, and cached renderer ($\mathrm{rad_gain} = 2.0$, compression 0.45, antialiasing 0.7 px, six meandering filaments with $w_0 = 0.50$ and $A_{\mathrm{MEANDER}} = 0.22$); fields are never re-sampled during optimization. The search vector is $\theta = (\kappa, \rho, \delta_p, \phi, \zeta)$ with bounds $\kappa \in \{0, \ldots, 5\}$, $\rho \in [0.5, 2.0]$, $\delta_p \in [-60, 60]$ px, $\phi \in [-15, 15]^\circ$, and $\zeta \in [0.5, 4.0]$ px. CMA-ES uses population 20, 50/25 iterations (1,000/500 queries per image for CLIP/LanguageBind versus IRGPT), and frozen per-image seeds starting at 20260901. Prompt strings, margins, and the deterministic 40-crop caption/VQA subset follow Section 3.6 and Appendix C.

For a compact comparison of the thermal-rendering controls, Table 1 reports the existing targeted top-1 ASR values on two representative IR scorers. Here and below, ASR denotes the targeted top-1 attack success rate. The table is deliberately limited to methods and outcomes; protocol and model details remain in the surrounding text and appendix.

Table 1: Comparison of thermal-rendering controls. Cells report existing targeted top-1 attack success rates under the IR-control protocol for the two representative scorers.

Method	Model	
	LanguageBind Thermal	IR-MetaCLIP-L14
Sensor drift	60.0%	23.0%
Atmospheric turbulence	48.3%	10.0%
Stripe noise	38.3%	18.0%
Structured plume (ours)	65.0%	60.0%

4.2 Primary Results: IR-Specialized Targets

Direct optimization is the central evaluation. LanguageBind reaches 67.1% (335/499) for person→vehicle and 27.3% (141/517) for vehicle→person. Under the frozen three-label IRGPT scorer, the target-specific top-1 rates are 19.9% (131/658) and 69.9% (327/468), respectively. The broader IRGPT open-set rates are 52.9% (348/658) and 82.9% (388/468), showing why target-specific and any-label flips must be reported separately.

Table 2: Core attack results on clean-correct subsets. Each cell reports only the target-specific top-1 flip rate. IRGPT open-set flips, which also count "other" outcomes, are reported in the text.

Target direction	Direct IR targets		IR-adapted encoders			
	LanguageBind Thermal	IRGPT (InternVL2-8B)	IR-OpenAI L14	IR-EVA G14	IR-MetaCLIP L14	IR-OpenCLIP B16
person→vehicle	67.1%	19.9%	<u>70.1%</u>	59.6%	71.8%	38.7%
vehicle→person	27.3%	69.9%	83.9%	63.6%	<u>83.5%</u>	80.7%

What does a successful attack look like? Figure 3 pairs clean-correct person crops with their plume-rendered top-1 flips. The paired rows keep the clean image, plume cue, and adversarial image aligned so that the target-adjacent change remains visible without presenting the field as a pasted texture. Open-ended classification, captioning, and VQA diagnostics are reported in Appendix D.

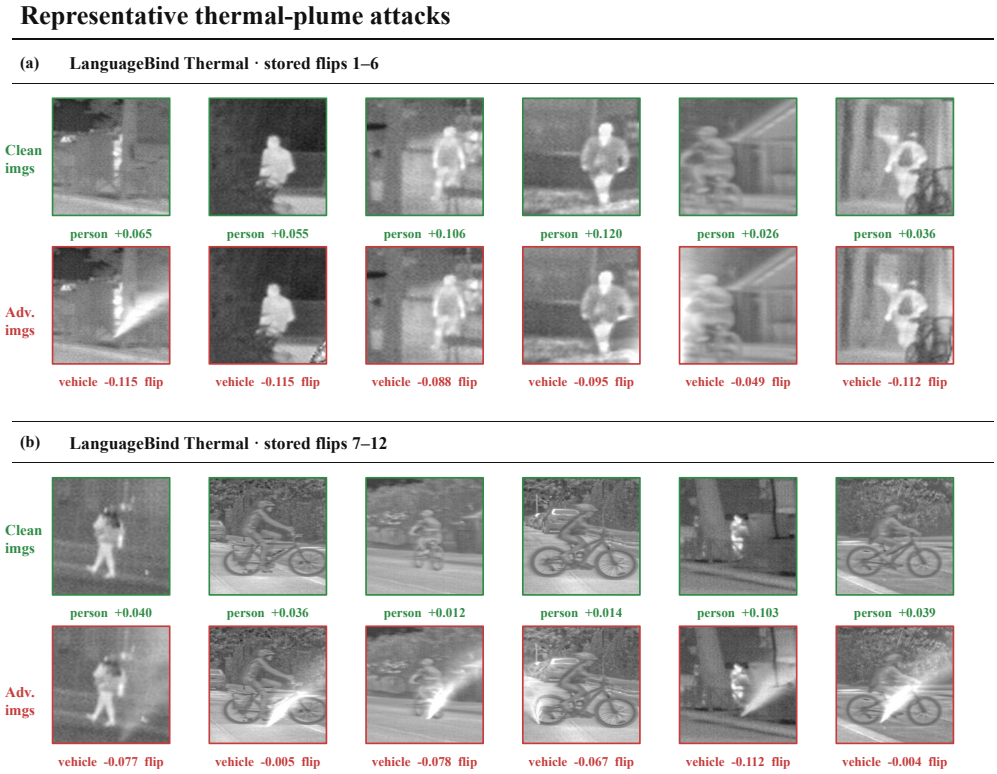

Figure 3: Representative thermal-plume attacks. Twelve stored pairs are arranged in two dense six-column panels: each column pairs a clean-correct crop (green) with its plume-rendered top-1 flip (red), with the predicted class and recorded LanguageBind margin shown below. Values are scorer margins, not softmax probabilities.

Do the changes persist beyond the closed-set label? We additionally probe one existing LanguageBind-optimized crop through the closed-set, VQA, and captioning evaluations. On this crop, IRGPT changes from "yes" to "no" while Qwen remains "yes"; the caption remains person-centered while changing its local heat description. Aggregate open-ended results are reported in Table 9.

4.3 IR-Adapted CLIP Variants

We next study four CLIP encoders adapted to infrared imagery. They form an adaptation analysis, not the primary IR-VLM evaluation, and use per-model prompts with $\alpha = 0.2$; direct IR targets and the secondary transfer diagnostic use frozen prompts with $\alpha = 0$. The variants are trained on synthetic infrared images with descriptive captions: OpenAI CLIP-L/14, EVA-CLIP-G/14, MetaCLIP-L/14, and OpenCLIP-B/16 (DataComp-XL). Their training distribution does not preserve the short-prompt geometry of the four-prompt baseline, leaving vehicle clean-correct rates at only 0.6–4.1% under those templates. We therefore select one person and one vehicle prompt on a separate 150-crop development split, maximizing the lower of the two gate rates, and freeze the choice before evaluation (prompts listed verbatim in Appendix C). The resulting person gates are 73.5–89.1% and vehicle gates 46.8–87.0%; every attacked image in Table 2 is clean-correct by construction, and each denominator equals the corresponding gate-pass count.

Table 2 gives the full attack on these clean-correct subsets. As in AirflowAttack (Su et al., 2026), we count a success when the top-1 label flips under the frozen per-model prompts. The attack flips 1,392/2,344 person windows (59.4%) and 988/1,302 vehicle windows (75.9%); these pooled totals aggregate different per-model gates and directions and are descriptive only, not a single-model ASR.

Across models, ASR spans 38.7–71.8% for person→vehicle and 63.6–83.9% for vehicle→person, with median doses of 0.5–1.3 px. The person rows span ADAS recordings plus LLVIP sequences, while the vehicle rows span ADAS recordings only.

What changes after adaptation? All four variants remain susceptible under the pre-specified protocol, with vehicle→person flips exceeding person→vehicle flips for every model. We report this directional asymmetry as an observation rather than attributing it to a particular plume mechanism. The three-class extension (Table 8) is model-dependent: OpenAI and MetaCLIP are near chance on "other", whereas EVA and OpenCLIP retain useful separation. We therefore retain binary top-1 ASR as the common adapted-model outcome and report the three-class extension separately.

4.4 Secondary Diagnostic: Transfer from Generic CLIP

Four generic encoders serve as frozen surrogate scorers in this secondary transfer diagnostic ($\alpha = 0$, $\approx$0.5 px dose). They reach 51.9–68.1% ASR on their own clean-correct decisions (Table 10), while the same per-image attacks flip only 3.9–8.4% of the matched LanguageBind overlap, versus 67.1% under direct LanguageBind optimization (Table 11). This gap motivates direct evaluation of the IR-specialized targets.

4.5 Matched Optimization Comparison

Table 3 compares optimization strategies on MetaCLIP under a matched evaluation setting. The structured plume reaches 60% ASR, compared with 28% for pixel-space CMA-ES, 26.7% for random sampling in the same five-dimensional space, and 4–23% for fixed controls. LPIPS and footprint differ across methods, so the result characterizes optimization under the stated protocol; the complete ASR and LPIPS records are given in Appendix F.

Table 3: Matched optimization comparison on MetaCLIP. LPIPS and footprint are not matched across controls.

Family	Variant	Targeted ASR
Proposed	**Structured plume (CMA-ES)**	**60.0%**
Search / parameterization	Pixel-space CMA-ES	28.0%
	Random sampling	26.7%
	Fourier	23.0%
	Affine	22.0%
Fixed controls	Random field	21.0%
	UAP	14.0%
	Noise	9.0%
	Patch	5.0%
	Blur	4.0%

4.6 Additional analyses

Stability and controls. Against the three listed synthetic approximations of IR phenomena, the plume reaches 65.0% on LanguageBind and 60% on MetaCLIP, exceeding every control in the matched comparison. Across the four IR-adapted encoders, the per-model standard deviation of ASR over five seeds (20260901–20260905), with 60 images per seed, is 1.4–2.4 percentage points (OpenAI L/14 2.4, MetaCLIP 2.2, EVA-G/14 1.9, OpenCLIP B/16 1.4). These stability runs use subsets and therefore do not replace the full-pool DEV estimates.

Holdout and open-set behavior. On the independent TEST-2 v2 holdout (60 unique LLVIP images across 26 sequences), LanguageBind ASR drops by 22.2 percentage points, from 67.1% to 44.9%. The expanded three-class protocol flips 48.2–61.0% of CLIP samples. In the open-ended VQA diagnostic, a clean "yes" becomes "no" on 27.0% of IRGPT samples (10/37 clean-yes) and 12.5% of Qwen samples (5/40). These answer spaces and denominators differ from the headline target-specific metric, so they are reported separately.

Table 4: Protocol-matched DEV and TEST-2 v2 attack success rates. The holdout is independent and uses a frozen single-run protocol.

Model	DEV	TEST-2 v2
LanguageBind Thermal	67.1%	44.9%
MetaCLIP L/14	68.1%	75.0%
OpenAI CLIP L/14	51.9%	28.8%
OpenCLIP B/16	58.8%	63.3%
EVA-G/14	55.0%	53.2%

Naturalness trade-off. Three stochastic automated-judge evaluations give a mean naturalness score of $3.5/5$. Raising α from 0 to 0.5 moves the ASR–LPIPS frontier from $80\%/0.138$ to $56\%/0.024$. Detailed tables, protocols, and the correction record are provided in Appendix H.

Physical-scale check. The chord-LOS anchors in Appendix A span 0.184–0.701 px for compact heaters and 2.51–4.61 px for surveillance-scale exhaust; these synthetic, uncalibrated values bracket the 0.5–4.0 px dose range used in the main experiments. A propane-heater distance sweep changes the anchor from 0.116 px at 4 m to 8.833 px at 0.25 m, documenting the attacker-distance degree of freedom rather than a hardware-calibrated claim.

Scope of the optimization. Each crop is optimized independently against its frozen scorer; no universal field is shared across images. The reported rates therefore measure per-image susceptibility under fixed query budgets, rather than transfer from a single universal perturbation.

5 Limitations

Thermal Plume Attack is a physics-inspired synthetic evaluation based on a reduced-order plume model with approximate radiance; it uses static 8-bit frames and does not model full radiative transfer, hardware calibration, temporal effects, or wavelength-dependent atmospheric absorption. Direct targets use scorer-specific query budgets and closed-set margin criteria, while open-ended VQA is reported separately, so all rates should be interpreted within the stated synthetic protocol.

6 Conclusion

Thermal Plume Attack is a physics-inspired synthetic benchmark for non-contact perturbations of infrared VLMs. For LanguageBind, target-specific top-1 ASR is 67.1% (335/499) for person→vehicle and 27.3% (141/517) for vehicle→person. IRGPT reaches 19.9% (131/658) and 69.9% (327/468) in the same two directions, while its broader open-set rates are 52.9% and 82.9%. The four IR-adapted encoders also remain susceptible under their stated protocol (38.7–71.8% and 63.6–83.9% for the two directions). An equal-budget, equal-α MetaCLIP comparison yields 60% ASR versus at most 28%, with LPIPS and footprint unmatched. Transfer from generic CLIP is far weaker (3.9–8.4% on the matched LanguageBind subset), so transfer diagnostics are not a substitute for direct optimization in this benchmark. All rates measure susceptibility under the synthetic protocol.

AI Use Statement

Generative AI tools were used only to aid or polish the writing of this paper, including language editing for clarity and grammar. They were not used to generate experimental data or numerical results, design the method or experiments, analyze results, or formulate mathematical claims. The authors reviewed the AI-assisted text and take responsibility for the final manuscript, figures, claims, and experimental artifacts.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Zhe Cao, Jin Zhang, and Ruiheng Zhang. Irgpt: Understanding real-world infrared image with bi-cross-modal curriculum on large-scale benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 166–176, 2025.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

Xiang Chen, Yingying Zhao, Chao Li, Jiaju Han, Ben Zhang, Ang Li, Jiahuan Long, Yiwei Wei, Jiujiang Guo, and Chengyin Hu. Qr-structured thermal triggers for targeted semantic attacks on infrared vision-language models. *arXiv preprint arXiv:2607.29445*, 2026.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 2206–2216, 2020.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Brandon Yushan Feng, Mingyang Xie, and Christopher A. Metzler. Turbugan: An adversarial learning approach to spatially-varying multiframe blind deconvolution with applications to imaging through turbulence. *IEEE Journal on Selected Areas in Information Theory*, 3(3):543–556, 2022.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Chengyin Hu, Yuxian Dong, Yikun Guo, Xiang Chen, Qike Zhang, Junqi Wu, Jiahuan Long, Jiujiang Guo, Yiwei Wei, Tingsong Jiang, and Wen Yao. Revealing physical-world semantic vulnerabilities: Universal adversarial patch for infrared vision-language models. *arXiv preprint arXiv:2604.03117*, 2026.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2137–2146, 2018.

Xin Li, Jiaju Han, Yaqi Ma, Chengyin Hu, Yingying Zhao, Jiahuan Long, Fengyu Zhang, and Yahui Chai. Infraqr: Edge-placed qr-inspired structured patch attacks on infrared vision-language models. *arXiv preprint arXiv:2607.07288*, 2026.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jiahuan Long, Tingsong Jiang, Hanqing Liu, Chao Ma, Weien Zhou, Yang Yang, and Wen Yao. Thermally activated dual-modal adversarial clothing against ai surveillance systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 34929–34939, 2026.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Zhiyuan Mao, Nicholas Chimitt, and Stanley H. Chan. Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14759–14768, 2021.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1765–1773, 2017.

Cong Su, Jiaju Han, Xuemeng Sun, Chengyin Hu, Qike Zhang, Jiujiang Guo, Yiwei Wei, and Jiahuan Long. Airflowattack: Thermal-airflow adversarial perturbations against infrared remote-sensing vision-language models. *arXiv preprint arXiv:2607.06485*, 2026.

Hui Wei, Zhixiang Wang, Xuemei Jia, Yinqiang Zheng, Hao Tang, Shin'ichi Satoh, and Zheng Wang. Hotcold block: Fooling thermal infrared detectors with a novel wearable design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15233–15241, 2023.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2739, 2019.

Yixing Yong, Jian Wang, Ming Lei, Lijun He, and Fan Li. Unleashing the representational power of fourier shapes for attacking infrared object detection. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 5005–5013, 2022.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:5642123, 2024.

Zhaolu Zheng, Gang Liu, Min Dang, and Jinpeng Luo. Black-box physical adversarial stripes for hiding from infrared detectors at multiple views. *Neural Networks*, 200:108821, 2026.

Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 6311–6320, 2023.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaopei Zhu, Zhanhao Hu, Siyuan Huang, Jianmin Li, and Xiaolin Hu. Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13317–13326, 2022.

A Rendering Details and Physical Anchors

The MTT mean field uses source diameter D, exit velocity V_0, and source temperature excess ΔT_0 to compute centerline velocity and temperature, followed by Gaussian radial profiles with

$$\sigma_u(z) = 0.089z, \qquad \sigma_T(z) = 0.128z. \tag{7}$$

The near-field mixing width is

$$\Delta T_{\text{near}}(z) = \max(\Delta T_{\text{MTT}}(z),\, 0.24D + 0.11z)\,. \tag{8}$$

Turbulence uses the Lane–Glindemann–Dainty FFT construction: white noise is filtered by $k^{-4/3}$ (power spectrum $k^{-8/3}$), scaled by $a = 0.085$, and clipped to $[T_{\text{amb}}, T_{\text{source}}]$. Gifford meandering uses

$$x_m(z) = A_{\text{MEANDER}}\sigma_T(z)\,\text{ramp}(z)s(z), \qquad A_{\text{MEANDER}} = 0.22, \tag{9}$$

with four sinusoidal modes spanning 1–3 m. Six independently meandering filaments use Gauss–Hermite anchors with $\sigma_c = 0.866\sigma_T$ and conserve variance via $\sigma_c^2 + \sigma_f^2 = \sigma_T^2$. Crosswind advection follows $dx_c/dz = u_{cw}/w(z)$.

Refraction uses

$$n - 1 = \frac{K_{\text{GD}}PM}{RT}, \qquad \Delta x = (n-1)\frac{f_{\text{px}}\text{LOS}_{\text{M}}}{\text{M}_{\text{PER_PX}}}, \tag{10}$$

with $f_{\text{px}} = 765$ for the FLIR configuration. The line-of-sight chord $\text{LOS}(z) = 2(0.35 + 3\sigma_T(z))$ is integrated along plume height. The `robust_dose` operation rescales the displacement field so its ROI-p95 equals the optimized dose. Radiance combines a white-hot overlay and temperature-field texture; the attack uses compression exponent 0.45, 0.7 px antialiasing, and $\text{rad_gain} = 2.0$. Table 5 lists the resulting displacement anchors.

Table 5: Physical displacement anchors (chord LOS, ROI-p95), computed from the physics-inspired model as synthetic, uncalibrated estimates.

Source	anchor (px)	Surveillance case	anchor (px)
Heat gun	0.184	Bus exhaust (trail 8 m, 2,000 px)	3.07
Car exhaust	0.233	Bus exhaust (trail 6 m, 3,000 px)	3.46
Propane heater	0.701	Bus exhaust (view 20 m bg, 2,000 px)	4.61
Space heater	0.196	Truck exhaust (6 m, 2,000 px)	2.51
Ground heat haze (20 m)	2.5	Ground heat haze (20 m, 2,000 px)	1.51

Table 6 isolates the distance dependence of the propane-heater anchor while holding its temperature excess and camera–target distance fixed.

Table 6: Attacker distance freedom for a propane heater ($\Delta T = 270$ K, camera–target distance 5 m).

Plume distance (m)	Anchor (px)	Relative
4.00	0.116	1.0×
3.00	0.310	2.7×
2.00	0.697	6.0×
1.00	1.860	16×
0.50	4.184	36×
0.25	8.833	76×

B Related-Work Comparison

Table 7 contrasts the target models, physical parameterization, and optimization settings of the two studies.

Table 7: Comparison with AirflowAttack (Su et al., 2026).

Dimension	AirflowAttack	**Thermal Plume Attack**
Target model	Generic CLIP for remote sensing	IR-specialized targets; generic CLIP references
Physical model	Heat-kernel airflow templates	MTT plume with turbulence, meander, filaments, refraction, and radiance
Optimization	Generator-based universal gray-box attack	Per-image score-only CMA-ES
Scenario	Remote-sensing classification	Person/vehicle crops from FLIR ADAS and LLVIP
Reported ASR	48.5% mean	19.9–69.9% target-specific ASR for direct IR targets; 51.9–68.1% for generic CLIP references

C Reproducibility Protocol, Prompts, and Scorers

All runs share the renderer, the crop pipeline, and the CMA-ES setup; only the scorer, the prompt set, α, and the query budget differ as specified below.

Preprocessing and rendering. Detector outputs (GT boxes) are padded by 40 px on every side to form the attack window, resized to 512×640 (bilinear), and converted to 8-bit apparent-temperature-like frames with the scene-adaptive gain `scene_gain_dn_per_k`. Rendering uses $\mathrm{rad_gain} = 2.0$, compression exponent 0.45, and 0.7 px antialiasing. The final plume fields use $N_{\mathrm{FIL}} = 6$ independently meandering filaments with width scale $w_0 = 0.50$ (relative meander $A_{\mathrm{MEANDER}} = 0.22$); fields are cached before evaluation and never re-sampled during optimization. CLIP-family inference runs in fp16; the final full-pool IRGPT directional matrix uses the explicitly recorded bf16 scoring path.

Optimizer. For every image, CMA-ES minimizes the scorer margin plus $\beta\zeta/d_{\max}$ with $\beta = 0.05$ and the optional α LPIPS term. The main-text notation is retained below:

- $\kappa \in \{0, \ldots, 5\}$ indexes the frozen condition fields below;
- $\rho \in [0.5, 2.0]$ is the plume scale;
- $\delta_p \in [-60, 60]$ px is the anchor offset from the target box;
- $\phi \in [-15, 15]^\circ$ is the plume lean angle;
- $\zeta \in [0.5, 4.0]$ px is the ROI-p95 displacement dose.

The optimizer starts from $\theta_0 = [\mathrm{uniform}(0, 6), 1.0, 0.0, 0.0, 1.0]$ with $\sigma_0 = 0.3$ and per-dimension scaling $[1.2, 0.3, 20, 5, 0.8]$; population 20, maxfevals $=$ pop $\times$ iters with iters $= 50$ for CLIP-class scorers (1,000 queries) and iters $= 25$ for the direct IRGPT run (500 queries). The frozen manifests and per-image CMA seeds start from 20260901 and advance by image index; the same seeds reproduce each manifest in the artifact.

Condition index. κ selects one of six frozen plume fields, each defined by a source, a crosswind condition, and an anchor position (P_c center, P_l left, P_r right of the target):

- 0: space heater, W0 crosswind, P_c;
- 1: heat gun (low), W0, P_c;
- 2: propane heater, W0, P_c;
- 3: space heater, W1, P_l;
- 4: heat gun (low), W2, P_r;
- 5: car-exhaust load, W0, P_c.

Each field is generated from the source's MTT parameters (source diameter, exit velocity, temperature excess; Table 5) with the shared turbulence, meander, and filament construction of Appendix A; the meander realization seed advances per condition, and the crosswind bend is advected kinematically.

Prompts (verbatim).

- LanguageBind Thermal (frozen gate pair): `a thermal image of a pedestrian` and `thermal vehicle`; the margin is the cosine similarity difference between the two.
- Forced-choice scorer (IRGPT direct runs, and the diagnostic caption protocol): `What is in this thermal image? Choose exactly one: person, vehicle, other.`
- IR-domain four-prompt sets (generic and IR-adapted runs before per-model selection), person: `a thermal infrared image of a person`, `a pedestrian in a thermal image`, `a human figure in infrared`, `a person captured by a thermal camera`; vehicle: `a thermal infrared image of a vehicle`, `a car, truck, or bus in a thermal image`, `a road vehicle in infrared`, `a vehicle captured by a thermal camera`.
- IR-adapted per-model dev-selected pairs (chosen on a disjoint 150-crop split, frozen before evaluation): IR-OpenAI-L14: `pedestrian`/`truck`; IR-EVA-G14: `pedestrian`/`a vehicle seen by a thermal camera`; IR-MetaCLIP-L14: `pedestrian`/`a car, truck, or bus in a thermal image`; IR-OpenCLIP-B16: `pedestrian`/`a car or truck`.
- Open-ended VQA: `Is there a person in this thermal image? Answer yes or no.`
- Captioning: `Describe this thermal image in one sentence.`
- Expanded three-class protocol: the "other" text scores come from a fixed candidate list (`other`, `something else`, `neither a person nor a vehicle`, `a scene without people or vehicles`, `background`, `nothing in particular`, `an empty scene`, `an object that is not a person or a vehicle`); the single best "other" prompt per model was chosen on the 150-crop dev split to maximize person/vehicle separation and is frozen in the artifact. The binary person–vehicle comparison is always evaluated separately and is the headline protocol.

Scorer and metric definitions.

- LanguageBind Thermal margin: $m = \cos(z_{\text{img}}, z_{\texttt{person}}) - \cos(z_{\text{img}}, z_{\texttt{vehicle}})$ with the gate-frozen text pair above; a crop is clean-correct iff $m > 0$; a flip is $m \leq 0$.
- IR-adapted and generic CLIP margin: $m = \frac{1}{4}\sum_i s_{\text{person},i} - \frac{1}{4}\sum_i s_{\text{vehicle},i}$ over the four prompt templates (or over the per-model pair when ptuned prompts are used).
- IRGPT sequence margin: with teacher-forced label log-probabilities $\log P(\ell)$ under the forced-choice prompt, the source-class margin is $m_s = \log P(s) - \max_{\ell \in \{\text{person,vehicle,other}\} \setminus \{s\}} \log P(\ell)$; clean-correct iff $m_s > 0$. A flip is recorded when the margin at the best-found attack is non-positive (the small $\beta\zeta/d_{\max}$ optimization regularizer only makes flips harder to record).
- VQA flip rate (Appendix D): $|\{\text{crop} : \text{clean} = \text{"yes"} \wedge \text{atk} = \text{"no"}\}| \,/\, |\{\text{crop} : \text{clean} = \text{"yes"}\}|$, with null answers counted as not-yes in a secondary view (none occurred).
- person-mention: fraction of generated captions containing "person"; ROUGE-L: LCS-based F1 between clean and attacked captions.

Model checkpoints. LanguageBind Thermal is the public LanguageBind "thermal" checkpoint; the IR-adapted encoders are LoRA/fully-fine-tuned variants of OpenAI CLIP L/14, EVA-G/14, MetaCLIP L/14, and OpenCLIP B/16 (DataComp-XL) trained on synthetic infrared crops with descriptive captions; IRGPT is the publicly released InternVL2-8B infrared checkpoint; Qwen2.5-VL-7B-Instruct is the public checkpoint. To make the evaluation reproducible, the supplementary package will include the fine-tuning recipes, frozen manifests, per-image results, scorer code, and the correction script referenced below.

D EXPANDED CLASSIFICATION AND MULTI-TASK RESULTS

Table 8: Three-class expanded prompt-set top-1 success. Best and second-best values are bold and underlined, respectively.

Scorer	Three-class top-1 ASR
MetaCLIP L/14	61.0%
OpenAI CLIP L/14	48.2%
OpenCLIP B/16	<u>55.5%</u>
EVA-G/14	52.9%

Protocol. The generative diagnostics evaluate plumes optimized for the LanguageBind scorer (1,000 queries/image, $\alpha = 0$) on the first 40 crops that are clean-correct under the LanguageBind gate ($m > 0$) in manifest order. These 40 crops are a deterministic subset of the DEV pool and are selected independently of the direct full-pool IRGPT runs; each crop is rendered with its own LanguageBind-optimized plume, and no additional optimization is run against the generative models, so the caption/VQA numbers below are not directly comparable to the direct optimization results in Table 2. IRGPT is scored by teacher-forced sequence log-probabilities over {yes, no} (deterministic, no decoding); Qwen2.5-VL-7B answers are decoded greedily and parsed by case-insensitive prefix or substring matching into yes/no labels (no null or conflicting answers occurred). IRGPT answered "yes" on 37/40 clean crops and Qwen on 40/40; these are the flip denominators. The VQA flip rate is $|\{\text{crop} : \text{clean} = \text{yes} \wedge \text{atk} = \text{no}\}|/|\{\text{crop} : \text{clean} = \text{yes}\}|$. person-mention is the fraction of generated captions containing the word "person"; ROUGE-L is the LCS-based F1 between the clean and attacked caption. IRGPT's caption proxy mode returns the forced-choice class instead of free-form captions, so for IRGPT "person-mention" counts crops whose forced-choice class is "person" and ROUGE-L is not applicable.

Table 9: Open-ended caption/VQA evaluation on the LanguageBind-optimized plumes. This is a stricter diagnostic than the headline margin flips of Table 2, with a different answer space and denominator. For IRGPT, "person-mention" is computed on the forced-choice output string (caption proxy) and ROUGE-L is not applicable.

	VQA	Caption diagnostic	
Model	yes→no	person-mention	ROUGE-L
IRGPT	27.0%	$0.925 \rightarrow 0.700$	n/a (proxy)
Qwen2.5-VL-7B	12.5%	$0.95 \rightarrow 0.90$	0.51

E SECONDARY TRANSFER REFERENCE

Table 10: Generic CLIP transfer-reference attacks. Best and second-best values are bold and underlined, respectively.

Scorer	Success rate
OpenAI CLIP L/14	51.9%
OpenCLIP B/16	<u>58.8%</u>
MetaCLIP L/14	68.1%
EVA-G/14	55.0%

Table 11: Transfer of surrogate-optimized plumes to the LanguageBind Thermal target on the matched clean-correct subset. Per-seed values are in the supplementary artifact. Best and second-best pooled ASRs are bold and underlined, respectively.

	LanguageBind Thermal target	
Surrogate	Pooled matched ASR	Per-seed range
OpenAI CLIP L/14	5.2%	[2.0, 8.0]
OpenCLIP B/16	8.4%	[5.5, 10.9]
MetaCLIP L/14	3.9%	[1.8, 5.4]
EVA-G/14	<u>5.7%</u>	[3.6, 8.9]

Multiple-comparison correction. Section 4.5 compares the plume against 20 baseline variants (two-proportion z-tests). We control the family-wise error rate with Bonferroni correction at level $\alpha/20$ and the false-discovery rate with the Benjamini–Hochberg procedure. The nine $\alpha = 0.2$ control comparisons in the recorded artifact remain significant after both corrections (all $p < 10^{-5}$). Across all 20 recorded comparisons, 13 survive Bonferroni correction and 14 survive Benjamini–Hochberg; the $\alpha = 0$ comparisons are therefore not uniformly significant. Because LPIPS and footprint are not matched in that comparison, these tests show that the plume differs from selected baselines under an equal query budget and equal α; they do not by themselves attribute the difference to plume structure (Section 4.5). Per-comparison p-values, effect sizes, and the correction script are included in the supplementary artifact file `multiple_comparison.json`.

F LPIPS Measurements and LanguageBind Baselines

We report LPIPS only for the configurations on which it was measured; other methods are intentionally omitted from the LPIPS record because their fixed-strength protocols do not produce the per-image LPIPS bookkeeping needed for a meaningful comparison. Table 12 lists all LPIPS measurements in this paper; a method absent from it has no LPIPS value to report, and no cell is left blank as a placeholder.

Table 12: All LPIPS measurements reported in this paper (methods not listed here had no LPIPS bookkeeping under their fixed-strength protocol). Plume configurations are shaded; no cross-metric ranking is implied.

	Measured outcome	
Configuration	ASR	LPIPS
Structured plume, α=0.2 (MetaCLIP)	60%	0.0626
Fourier, α=0.2	23%	0.0077
UAP, α=0.2	14%	0.0044
Patch, α=0.2	5%	0.0053
Plume, α=0 (ASR–LPIPS frontier)	80%	0.138
Plume, α=0.5 (ASR–LPIPS frontier)	56%	0.024

Table 13: Matched optimization comparison on MetaCLIP. LPIPS and footprint are not matched across methods; available LPIPS values are in Table 12. The proposed method is shaded; best and second-best ASRs are bold and underlined, respectively.

	Measured outcome
Method	ASR
Structured plume (CMA-ES)	60%
Random sampling	26.7%
Pixel-space CMA-ES	<u>28%</u>
Fourier	23%
Affine	22%
Random field	21%
UAP	14%
Noise	9%
Patch	5%
Blur	4%

Table 14 lists all generic control baselines on the LanguageBind scorer at $\alpha = 0$, under the same protocol as the plume's direct LanguageBind run. Without the LPIPS regularizer unstructured baselines can exceed the plume. This mirrors the finding on the CLIP side that unconstrained perturbation strength dominates when constraints are removed (Section 4.5); the IR-specific baselines in Table 1, by contrast, all remain below the plume even at $\alpha = 0$. Table 13 shows the plume under an explicit LPIPS penalty at $\alpha = 0.2$; as stated there, LPIPS and footprint are not matched across methods, so none of these tables should be read as evidence that plume structure alone outperforms controls.

Table 14: Generic control baselines on the LanguageBind scorer under the same protocol as the direct plume run. The proposed method is shaded; the best and second-best values are bold and underlined, respectively.

Method	LanguageBind ASR
Random field	95.0%
Affine	<u>78.3%</u>
Structured plume (ours)	65.0%
Blur	56.7%
Patch	38.3%
Noise	33.3%
Brightness	30.0%
Fourier	10.0%
UAP	5.0%

G TEST-2 HOLDOUT DETAIL

TEST-2 v2 is an independent, frozen holdout of 60 unique LLVIP images, one window per image, stratified across 26 acquisition sequences. It is disjoint from DEV and the gate TEST set. The earlier v1 split drew 60 crops from only 34 unique images, all from a single LLVIP sequence, and is excluded from the comparison. Table 15 reports the protocol-matched DEV versus TEST-2 v2 comparison for LanguageBind and the generic CLIP references. LanguageBind drops from 67.1% to 44.9%, while the CLIP encoders show heterogeneous behavior (MetaCLIP 75.0%, OpenCLIP B/16 63.3%, EVA-G/14 53.2%, OpenAI CLIP 28.8% on TEST-2). The earlier IRGPT holdout artifact was produced by a superseded scoring path and is therefore excluded rather than mixed with the final full-pool bf16 directional matrix. These results support split-local efficacy; robust OOD generalization is not established.

Table 15: TEST-2 v2 holdout under the frozen single-run protocol. The superseded IRGPT holdout artifact is excluded; the v1 split is also excluded (see text).

	Attack success rate	
Model	DEV	TEST-2 v2
IR-specialized targets		
LanguageBind Thermal	67.1%	44.9%
Generic CLIP references		
MetaCLIP L/14	68.1%	75.0%
OpenAI CLIP L/14	51.9%	28.8%
OpenCLIP B/16	58.8%	63.3%
EVA-G/14	55.0%	53.2%

H Additional Analyses

Synthetic IR controls and optimization stability. Against the three listed synthetic approximations of IR phenomena, the plume reaches 65.0% on LanguageBind and 60% on MetaCLIP, exceeding every control in the matched comparison of Table 1. For the four encoders, the per-model standard deviation of ASR across five seeds (20260901–20260905), each evaluated on 60 images per seed, is 1.4–2.4 percentage points (OpenAI L/14 2.4, MetaCLIP 2.2, EVA-G/14 1.9, OpenCLIP B/16 1.4); these stability runs use 60-image subsets, so their point estimates need not equal the full-pool DEV rates.

Holdout and open-set behavior. On the independent TEST-2 v2 holdout (frozen, disjoint, 60 unique LLVIP images across 26 sequences), LanguageBind ASR drops by 22.2 percentage points, supporting split-local rather than robust OOD efficacy (Section G). The expanded three-class protocol flips 48.2–61.0% of CLIP samples. Under the open-ended VQA diagnostic, a clean "yes" becomes "no" on 27.0% of IRGPT samples (10/37 clean-yes) and 12.5% of Qwen samples (5/40) (Section D). This diagnostic uses a different answer space and denominator from the headline target-specific metric and must not be read as the same quantity.

Naturalness. Three stochastic automated-judge evaluations (an LLM with vision, three temperature/seed variants) give a mean naturalness score of $3.5/5$; raising α from 0 to 0.5 moves the ASR–LPIPS frontier from $80\%/0.138$ to $56\%/0.024$ (Section F).

Ethics and data. All experiments use public FLIR ADAS and LLVIP data and synthetic thermal patterns parameterized by everyday heat sources. The study involves no human subjects or private data.