Avaliação da Percepção de Segurança Urbana no Distrito Federal utilizando Grandes Modelos Multimodais e Imagens de *Street View*

Lucas Adeodato[1], Pedro Nascimento[2], Danilo Melo[2],
Thiago Lourenço[2], Lucas de Oliveira[2] e Gustavo Dias[1]
[1]Universidade de Brasília (UnB), Brasília/DF - Brasil
[2]Deep Dive, Brasília/DF - Brasil

***Resumo*— A percepção de segurança urbana é insumo do planejamento de cidades inteligentes, e medi-la por pesquisa de opinião é caro e pouco escalável. Este trabalho descreve uma metodologia de análise estruturada para mapear essa percepção no Distrito Federal a partir de imagens no nível da rua (SVI) julgadas por Grandes Modelos Multimodais de Linguagem (MLLMs), com validação prevista contra um padrão de referência humano coletado por pesquisa de opinião. A campanha com respondentes segue em campo, e reportamos aqui a caracterização das rodadas de modelo que a antecedem: seis rodadas de comparação pareada com MLLMs de pesos abertos executados localmente, combinando três modelos e duas formulações de *prompt* sobre 1.000 localizações, com 14.000 confrontos cada. Para cada rodada estimamos quanto do ranqueamento é sinal e quanto é ruído de amostragem, e usamos essa estimativa como teto ao comparar as rodadas entre si. A confiabilidade varia de 0,92 a 0,19 sob desenho idêntico, e cai sob a formulação mais longa apenas na família Qwen3-VL. Trocar a pergunta desloca até 39% do sinal estável de um mesmo modelo, e dois modelos sob a mesma pergunta compartilham cerca de metade da variância do ranqueamento. *Prompt* e modelo são, portanto, fatores experimentais que qualquer validação humano-IA precisa declarar.**

***Palavras-Chave*— Percepção de Segurança Urbana, Grandes Modelos Multimodais, *Street View Images*.**

I. Introdução

A segurança urbana condiciona a qualidade de vida nas cidades. A percepção de segurança não decorre apenas de índices objetivos de criminalidade: ela responde também a fatores do ambiente construído, como iluminação, manutenção de vias, presença de vegetação e movimento de pessoas [1], [2]. Tradicionalmente, essa percepção é medida por pesquisas de opinião, o que é custoso e de difícil escalabilidade.

Com o avanço das tecnologias de mapeamento, como as Imagens no Nível da Rua (*Street View Images*, SVI), tornou-se possível analisar o ambiente urbano em larga escala, e trabalhos pioneiros como o do Instituto de Tecnologia de Massachusetts (MIT) [3] demonstraram a viabilidade de usá-las para prever percepção de segurança. A chegada dos Grandes Modelos Multimodais de Linguagem (MLLMs) acrescentou a possibilidade de avaliar a cena sem treinamento supervisionado prévio [4].

Este trabalho propõe uma metodologia de análise comparativa para avaliar e mapear a percepção de segurança urbana no Distrito Federal (DF), Brasil, a partir de SVIs julgadas por MLLMs. A metodologia adapta o trabalho de Zhang *et al.* [4] e acrescenta a validação cruzada contra um padrão de referência humano coletado via pesquisa de opinião. A contribuição desta etapa é metodológica e vem acompanhada de resultados preliminares do lado dos modelos: a sequência de etapas completa, da amostragem espacial ancorada na malha viária até a agregação dos votos pareados, bem como a caracterização empírica de seis rodadas de MLLM, na Seção IV. Essa caracterização responde às perguntas que precedem qualquer confronto com o juízo humano: quanto de cada ranqueamento sobreviveria a repetir a rodada, quanto ele se desloca ao se reformular a pergunta e quanto dois modelos concordam entre si. A campanha com respondentes humanos segue em campo, e a comparação humano-IA fica para a etapa seguinte.

II. Trabalhos Relacionados

A interseção entre inteligência artificial e avaliação do ambiente construído evoluiu por camadas, que organizamos em quatro eixos: a fundamentação teórica da percepção de segurança; a medição por comparação pareada e sua agregação estatística; a predição em larga escala por visão computacional; e a integração recente de MLLMs para avaliação direta, sem exemplos prévios (*zero-shot*).

A. Percepção de segurança e o ambiente construído

A hipótese de que atributos físicos do espaço urbano influenciam tanto o comportamento criminal quanto a sensação de segurança remonta ao *Crime Prevention Through Environmental Design* (CPTED) e à noção de espaço defensável de Newman [5], que articula vigilância natural, controle de acesso, territorialidade e manutenção como dimensões projetáveis. Esses fatores ajudam a explicar por que um lugar transmite (in)segurança. A percepção que o pedestre forma diante de uma cena, porém, é um juízo holístico e instintivo, e é esse juízo global que este trabalho busca medir, sem decompô-lo em critérios. A ligação empírica entre forma urbana e segurança foi reforçada por Abraham *et al.* [1], que correlacionaram espacialmente a "Segurança Visualmente Percebida" e ocorrências criminais em Estocolmo, e por Xie *et al.* [2], que isolaram por SHAP (SHapley Additive exPlanations) quais elementos visuais da rua mais contribuem para a modelagem de criminalidade.

B. Medição por comparação pareada e agregação

A medição quantitativa da percepção urbana em escala foi consolidada pelo *Place Pulse* 1.0 [6], que agregou

comparações pareadas de imagens de rua ("qual lugar parece mais seguro?") em escalas perceptuais consistentes entre cidades. O *Place Pulse* 2.0 [7] ampliou a coleta para 56 cidades e converteu as comparações em escores pelo algoritmo *TrueSkill* [8]. A opção pela comparação pareada, em lugar de notas absolutas, apoia-se na maior consistência dos julgamentos relativos entre respondentes. Para transformar votos pareados em ranqueamento, além do *TrueSkill*, o modelo de Bradley-Terry [9] é o estimador clássico, com inferência eficiente viabilizada por algoritmos MM [10] e pela versão espectral ILSR [11], robusta a grafos esparsos.

C. Predição em larga escala: de regressores supervisionados a VLMs

Historicamente, extrair percepções de SVIs dependia de atributos manuais ou de Redes Neurais Convolucionais supervisionadas. O trabalho fundacional de Naik *et al.* [3] (*Streetscore*) treinou um regressor sobre os rótulos do *Place Pulse* e previu segurança percebida em um milhão de cenas, ao custo de exigir vastos conjuntos anotados, o que limitava a transferência para novos contextos geográficos ou culturais. Modelos de visão-linguagem (Vision-Language Models - VLMs) com pré-treino contrastivo (CLIP, do inglês Contrastive Language-Image Pretraining) [12] e, depois, SigLIP [13], que troca a perda *softmax* por uma perda sigmoide pareada) passaram a fornecer representações transferíveis sem rótulos específicos da tarefa, e o UrbanCLIP [14] demonstrou o potencial desse alinhamento para o perfilamento de regiões urbanas. No contexto latino-americano, Moreno-Vera *et al.* [15] aplicaram aprendizado profundo a SVIs para quantificar percepção de segurança. Estudos de cobertura alertam, contudo, que a disponibilidade e a idade das imagens do *Street View* na América Latina são limitadas [16], o que motiva o filtro temporal (2024 em diante) e a amostragem ancorada na malha viária adotados aqui.

D. MLLMs para avaliação zero-shot da segurança urbana

Zhang *et al.* [4] introduziram uma metodologia na qual MLLMs como o GPT-4o avaliam a segurança urbana de forma *zero-shot*, e extrapolam os rótulos gerados em um conjunto âncora para toda a cidade via características CLIP e o algoritmo de k-vizinhos mais próximos (K-NN), o que contém o custo de inferência em larga escala. A aderência desses modelos à percepção humana tem sido objeto de *benchmarks* dedicados: Mushkani [17] observou que VLMs alinham-se melhor com humanos em atributos visualmente objetivos do que em julgamentos subjetivos; Lan [18] reportou cerca de 89% de concordância entre um sistema VLM e a percepção agregada de moradores locais, resultado que sustenta o valor de validar a IA contra a população. A limitação de iluminação das SVIs diurnas tem sido endereçada por geração dia-para-noite com IA generativa [19], direção reservada para trabalhos futuros.

Este trabalho parte dessas contribuições e se distingue delas em dois pontos. O primeiro é não tratar a saída do MLLM como padrão de referência: propomos a coleta de uma linha de base humana local, agregada por Bradley-Terry, para medir a aderência das IAs multimodais à realidade percebida pela população do DF. O segundo é tratar o modelo e a formulação do *prompt* como fatores experimentais, e não como detalhes de implementação, já que ambos deslocam o ranqueamento o bastante para condicionar a leitura de qualquer métrica de concordância. A extrapolação posterior dos escores para toda a cidade, por representações contrastivas combinadas a K-NN, no espírito de Zhang *et al.* [4], é etapa subsequente.

III. Metodologia

A Figura 1 detalha o fluxo de trabalho proposto, estruturado em quatro macroetapas: a amostragem espacial e a coleta de imagens no nível da rua; a estruturação de um padrão de referência humano por meio de uma pesquisa de opinião; a inferência *zero-shot* com MLLMs; e a medição da concordância entre o juízo humano e as escolhas da inteligência artificial.

A. Amostragem e Coleta de Imagens

Foi desenvolvida uma infraestrutura para extração de SVIs da malha viária do DF. A amostragem dos pontos é proporcional à densidade viária, de modo a refletir a "experiência média" da população e as áreas de maior tráfego, e prioriza imagens a partir de 2024. A coleta abrange quatro vistas cardinais (0º, 90º, 180º, 270º) por ponto. A Figura 4 mostra a distribuição espacial resultante.

O conjunto de avaliação usado nas rodadas da Seção IV é um conjunto âncora de 1.000 localizações, com uma única imagem por localização, fixada na orientação de 0º. Manter as quatro vistas dentro de um mesmo ranqueamento misturaria variação de lugar com variação de ângulo: em rodadas preliminares sobre 250 localizações com as quatro vistas, três quartos da variância do escore de um dos modelos estava entre vistas do mesmo ponto. Uma imagem por localização elimina essa confusão e quadruplica o *n* espacial.

B. Coleta de Padrão de Referência Humano

Para validar a precisão dos MLLMs no contexto brasileiro e contornar vieses dos modelos fundacionais, desenvolvemos uma plataforma de pesquisa baseada no protocolo pareado do *Place Pulse* 2.0 [7]. A coleta humana opera sobre o mesmo conjunto âncora de 1.000 imagens e foca em um único eixo: a cada rodada, o respondente faz uma escolha binária forçada para a pergunta "qual lugar parece mais seguro?". Nenhum critério objetivo lhe é fornecido, diferentemente de Zhang *et al.* [4], que apresentam uma lista de atributos; o objetivo é capturar o juízo instintivo da cena. Cada sessão tem 35 comparações, com a posição esquerda/direita randomizada no servidor para neutralizar viés posicional.

Os votos das sessões completas são agregados por um modelo de Bradley-Terry [9], em que a probabilidade de uma imagem i ser preferida a j é $\sigma(\beta_i - \beta_j)$. A estimação utiliza uma iteração espectral conforme ILSR [11], com uma distribuição *a priori* regularizadora que mantém o ajuste identificável em grafos desconexos, comuns no início da coleta, e o ranqueamento é recalculado em lote de forma assíncrona. A amostragem de pares emprega aprendizado ativo leve: prioriza imagens sub-amostradas e, em parte das rodadas, escolhe o par mais informativo segundo a proximidade dos escores β. A campanha segue em andamento e a linha de base humana não é reportada nesta versão.

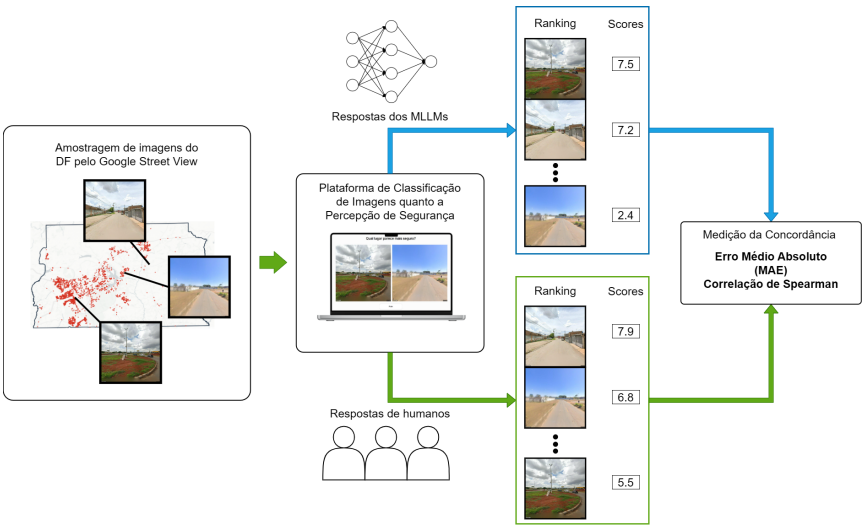


Fig. 1. Diagrama metodológico: coleta de imagens, estruturação do padrão de referência humano, inferência via MLLMs e avaliação da concordância.

C. Comparações pareadas pelos MLLMs

Avaliamos MLLMs de pesos abertos, executados localmente em suas variantes *instruct* por um ponto de acesso compatível com a API da OpenAI: Gemma 3 12B (Google DeepMind) [20] e Qwen3-VL nos tamanhos 4B e 8B (Alibaba Cloud) [21]. A execução local zera o custo marginal por confronto, o que viabiliza repetir a rodada a cada mudança de fator, e fixa a versão do modelo. Cada MLLM é submetido ao mesmo protocolo pareado dos humanos: dado um par de imagens, o modelo responde "A" ou "B". Não há opção de empate, o que força a discriminação também nos pares próximos, e a temperatura é fixada em 0,1.

Duas formulações de *prompt* são contrastadas, com modelo e conjunto de imagens fixos. A **específica** ancora o julgamento em um sujeito e em um contexto temporal (*"Which place feels safer for a pedestrian walking alone at night? Type only 'A' for the left image or 'B' for the right."*); a **genérica** remove esse ancoramento e pergunta apenas *"Which place feels safer?"*, preservando o restante da instrução. Os três modelos foram executados sob as duas, o que fecha um desenho fatorial de seis rodadas e isola o efeito da formulação sobre o ranqueamento.

As pontuações das rodadas de IA são obtidas por atualização Elo incremental ($K = 24$, inicialização em 1500) ao longo da execução, enquanto a linha de base humana é agregada por Bradley-Terry. Os dois estimadores produzem escalas distintas e não diretamente comparáveis. Por isso toda comparação neste trabalho, entre rodadas e futuramente contra o juízo humano, é feita em posto, por correlação de Spearman, que é invariante a transformações monótonas da escala.

D. Confiabilidade e teto de atenuação

Cada imagem i participa de n_i confrontos e vence w_i; a taxa observada $\hat{p}_i = w_i/n_i$ é uma estimativa ruidosa de sua força latente. A variância entre imagens decompõe-se em duas parcelas, sinal e ruído binomial de amostragem:

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}(p) + \mathrm{E}\left[\frac{p(1-p)}{n}\right] \qquad (1)$$

Definimos a **confiabilidade** da rodada como $r = \mathrm{Var}(p)/\mathrm{Var}(\hat{p})$, a fração da variação do ranqueamento que sobreviveria a repetir a rodada com o mesmo modelo. Essa quantidade é o teto da correlação que a rodada pode ter com qualquer outra medida: duas rodadas independentes com confiabilidades r_A e r_B não podem correlacionar acima de $\sqrt{r_A \cdot r_B}$, ainda que meçam exatamente a mesma grandeza. Reportamos, para cada par de rodadas, o ρ de Spearman observado e a razão $\rho/\sqrt{r_A \cdot r_B}$, que separa discordância substantiva de atenuação por ruído. Duas aproximações são assumidas: a taxa de vitória ignora a força do adversário, o que com pareamento aleatório sobre 1.000 itens acrescenta ruído pequeno, e a decomposição usa a taxa de vitória em vez do Elo, embora os dois correlacionem a $\rho \geq 0,96$ dentro de cada rodada.

Métricas de desvio absoluto, como Erro Absoluto Médio e R^2, exigiriam reestimar as rodadas de IA sob o estimador do padrão de referência humano, o que depende da preservação dos votos brutos discutida na Seção V.

TABELA I

CONFIABILIDADE DAS SEIS RODADAS DE MLLM

Modelo	*Prompt*	dp obs.	dp ruído	r
Gemma 3 12B	específico	0,283	0,081	0,92
Gemma 3 12B	genérico	0,284	0,081	0,92
Qwen3-VL 4B	específico	0,107	0,096	0,19
Qwen3-VL 4B	genérico	0,182	0,091	0,75
Qwen3-VL 8B	específico	0,178	0,091	0,74
Qwen3-VL 8B	genérico	0,262	0,083	0,90

Desvio-padrão observado e do ruído binomial da taxa de vitória por imagem; r é a confiabilidade da Seção III-D.

IV. Resultados e Discussão

A. Desenho das rodadas e confiabilidade

Foram executadas seis rodadas sobre as mesmas 1.000 imagens, com 14.000 confrontos cada e 28 aparições por imagem em média (mínimo 12, máximo 47). O desenho é idêntico entre elas; variam o modelo e a formulação do *prompt*, como resume a Tabela I.

As rodadas não são intercambiáveis em qualidade. Com 28 confrontos por imagem, um modelo que escolhesse ao acaso ainda produziria dispersão na taxa de vitória de aproximadamente $0,5/\sqrt{28} \approx 0,094$. O Qwen3-VL 4B sob a formulação específica tem dispersão observada de 0,107, ou seja, quase

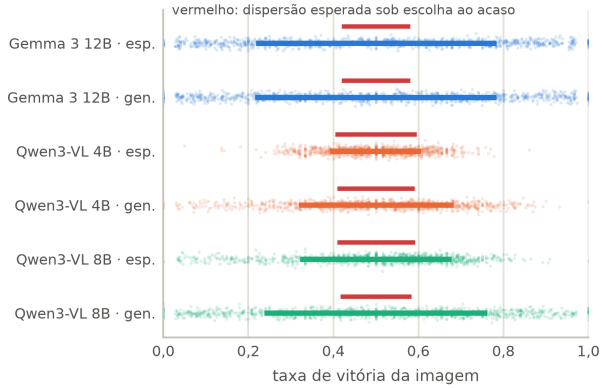


Fig. 2. Dispersão da taxa de vitória por imagem. O traço colorido marca ± 1 desvio-padrão observado e o vermelho, ± 1 desvio-padrão do ruído binomial esperado sob escolha ao acaso. Faixas próximas indicam pouco sinal no ranqueamento.

TABELA II

CONCORDÂNCIA ENTRE PARES DE RODADAS

Par de rodadas	ρ	**teto**	**ρ/teto**
Mesmo modelo, formulações distintas			
Gemma 3 12B	0,80	0,92	0,87
Qwen3-VL 4B	0,14	0,38	0,37
Qwen3-VL 8B	0,50	0,81	0,61
Mesma formulação, modelos distintos			
Gemma 3 12B × Qwen3-VL 4B (gen.)	0,65	0,83	0,78
Gemma 3 12B × Qwen3-VL 8B (gen.)	0,65	0,91	0,71
Qwen3-VL 4B × Qwen3-VL 8B (gen.)	0,70	0,82	0,86
Gemma 3 12B × Qwen3-VL 8B (esp.)	0,49	0,82	0,59

ρ de Spearman sobre o escore Elo; teto de atenuação $\sqrt{r_A r_B}$. Omitem-se os pares entre modelos com a rodada Qwen3-VL 4B específica, não interpretáveis com $r = 0,19$.

toda a variação do seu ranqueamento cabe dentro do que o acaso explica. Na Figura 2, a faixa observada dessa rodada mal excede a esperada sob escolha aleatória, enquanto nas demais ela é de duas a três vezes maior.

A perda de confiabilidade sob a formulação específica acompanha o modelo, não a dificuldade da pergunta. Ela ocorre nas duas rodadas Qwen3-VL, de 0,75 para 0,19 no 4B e de 0,90 para 0,74 no 8B, e em nenhuma das duas do Gemma 3 12B, que fica em 0,92 sob ambas. Se a causa fosse o esforço adicional que a pergunta mais longa exige, o Gemma também cairia. A hipótese mecanicista é a interação entre o orçamento de *tokens* da requisição e o raciocínio explícito, que o Qwen3-VL usa e o Gemma 3 não: consumido o orçamento dentro do bloco de raciocínio, o campo de conteúdo retorna vazio e o analisador sintático extrai um "A" ou "B" solto do raciocínio truncado. A formulação específica é mais longa e induz mais raciocínio, o que torna a truncagem mais provável. Confirmar a hipótese exige re-executar as rodadas com orçamento ampliado e registrar o motivo de encerramento de cada resposta. Independentemente da causa, o diagnóstico é aqui um controle de qualidade: sem ele, a ordenação da rodada de $r = 0,19$ seria indistinguível de uma legítima.

B. Sensibilidade à formulação do prompt

Com o desenho fechado, a troca de pergunta pode ser medida dentro de cada um dos três modelos. A Tabela II reúne esses pares e os pares entre modelos.

O efeito da formulação existe e não é uniforme entre modelos. No Gemma 3 12B, trocar a pergunta produz $\rho = 0,80$ contra um teto de 0,92, razão de 0,87: cerca de 13% do sinal estável do ranqueamento depende de qual pergunta foi feita, e 48 das 100 imagens mais seguras são as mesmas nas duas variantes. No Qwen3-VL 8B a razão cai para 0,61, com quase 40% do sinal estável dependendo da formulação, e as duas rodadas são confiáveis o bastante (0,90 e 0,74) para que isso não seja atribuível a ruído. Esse par mostra ainda onde a reordenação se concentra: apenas 15 das 100 imagens mais seguras coincidem entre as formulações, contra 52 das 100 menos seguras. Perguntar por onde se andaria sozinho à noite reescreve a ponta segura e deixa a insegura quase intacta. No Qwen3-VL 4B, $\rho = 0,14$ contra um teto de 0,38 mede a rodada de confiabilidade 0,19, não a sensibilidade ao *prompt*.

A consequência de desenho é direta: qualquer número de concordância entre modelos, ou contra o padrão de referência humano, precisa declarar sob qual formulação foi obtido.

C. Concordância entre modelos

Sob a formulação genérica há três rodadas comparáveis, o que separa duas perguntas: se modelos de famílias distintas concordam, e se aumentar o número de parâmetros altera o juízo. Os modelos concordam parcialmente, e a discordância que resta é substantiva: mesmo descontada toda a atenuação, dois deles ordenando as mesmas 1.000 imagens sob a mesma pergunta compartilham cerca de metade da variância. É uma melhora expressiva sobre as rodadas preliminares com as quatro vistas cardinais, em que ρ era 0,13. Ainda assim, o resultado é incompatível com tratar "o MLLM" no singular: a escolha do modelo altera o ranqueamento na mesma ordem de grandeza do efeito que se quer medir. O par de tamanho (Qwen3-VL 4B contra 8B, razão 0,86) é o mais concordante do conjunto, e o salto de 4B para 8B aparece sobretudo na confiabilidade, que vai de 0,75 para 0,90.

Sob a formulação específica, o par Gemma 3 12B contra Qwen3-VL 8B cai de razão 0,71 para 0,59: perguntar pelo pedestre sozinho à noite custa confiabilidade e ainda afasta os modelos entre si. O par mais distante do conjunto combina as duas trocas, com razão de 0,29 entre o Gemma 3 12B genérico e o Qwen3-VL 8B específico.

D. Como cada rodada ordena as cenas

As seções anteriores tratam de posto e correlação. A Figura 3 mostra as imagens, e nela aparecem dois critérios visuais distintos entre as rodadas confiáveis, em vez de um só.

Quatro rodadas (Gemma 3 12B nas duas formulações, Qwen3-VL 4B e 8B na genérica) repetem o mesmo repertório. No extremo seguro: via asfaltada com meio-fio, edificação de fachada mantida atrás de muro baixo ou grade, vegetação tratada. No inseguro: via de terra, muro cego ou manchado, lote vago, obra parada. O Qwen3-VL 8B sob a formulação específica adota outro critério, com confiabilidade de 0,74: no topo estão ruas comerciais com fachada ativa, carro estacionado e gente na calçada; no fundo, terreno vazio, campo aberto e terraplenagem, incluindo uma via bem asfaltada e deserta que as outras rodadas colocariam entre as melhores. É a leitura que a pergunta pede, já que quem anda sozinho à noite avalia menos o acabamento da fachada e mais se há alguém por perto. O Qwen3-VL 4B específico não sustenta



Fig. 3. As cinco imagens de escore Elo mais alto e mais baixo em cada rodada. Quatro rodadas ordenam por acabamento da via e da fachada; a Qwen3-VL 8B específica ordena por fachada ativa e movimento. Onde há empate no extremo, a seleção segue o Elo, que desempata pela força do adversário sorteado.

critério algum: via de terra e mato alto ocupam as duas pontas, comportamento previsto por uma confiabilidade de 0,19.

Duas ressalvas acompanham a leitura. A formulação que produziu o critério distinto no Qwen3-VL 8B também lhe custou 16 pontos de confiabilidade; se a causa for a truncagem descrita acima, o que sobrou não é uma subamostra aleatória dos juízos do modelo. E a imagem individual não é estável em rodada alguma: nos 15 pares possíveis, a interseção das cinco melhores é não vazia em dois, e a das cinco piores em três, sempre com uma imagem em comum.

E. Associação preliminar com criminalidade registrada

Como validação externa independente da linha de base humana, cruzamos os escores com microdados criminais de 2024 a 2025 obtidos junto à Polícia Civil do DF por Lei de Acesso à Informação (149.623 eventos geolocalizados), agregados em células H3 nas resoluções 7, 8 e 9. Para cada célula ajustamos um modelo de Poisson sobre a contagem de crimes contra o pedestre, com $\log$(km de via) como termo de compensação, $\log$(população + 1) do Censo 2022 como covariável e erros-padrão agrupados por Região Administrativa, já que células vizinhas não são independentes. O parâmetro de interesse, β_1, é o efeito de um desvio-padrão de escore perceptual sobre o log-crime.

Onde há coeficiente distinto de zero, ele é positivo: célula percebida como mais segura tem *mais* crime contra pedestre por km de via. O sinal é contraintuitivo, e a explicação plausível é exposição, já que lugar de aparência segura tende a ser urbanizado e com fluxo de pedestres. O efeito, porém, não sobrevive a duas verificações. Ele encolhe com a resolução: no Gemma 3 12B específico, β_1 cai de $+0,47$ em resolução 7 para $+0,09$ em resolução 9, padrão típico de artefato de agregação. E parte da geocodificação da base criminal ancora registros no centroide do setor censitário: 56 das 494 células em resolução 8 (11,3%) concentram 37,4% do crime contra pedestre, e excluindo-as todos os coeficientes caem para dentro do intervalo de confiança do zero. A rodada que mais se aproxima de escapar é o Qwen3-VL 8B específico ($+0,43$, erro-padrão 0,16 em resolução 8; $+0,26$, erro-padrão 0,15, $p = 0,09$ após o filtro).

A leitura adequada é que as rodadas avaliadas não produzem evidência de associação entre escore perceptual e crime registrado, o que também não estabelece ausência de associação: com 494 células e contagem sobredispersa, apenas efeitos moderados a grandes seriam detectáveis, e o desenho depende de crime *registrado*, que mistura ocorrência com propensão a notificar.

F. Uma restrição de desenho para a validação humana

Os extremos dos ranqueamentos não são aleatórios no espaço, como mostra a Figura 4. Nas quatro rodadas do primeiro critério visual, as 50 melhores concentram-se em Brasília, Lago Sul, Park Way e Sudoeste, e as 50 piores em Ceilândia, Samambaia, Recanto das Emas e Estrutural, com fração urbana mediana de 1,00 nas duas pontas. O que elas reproduzem é a ordenação de renda entre as Regiões Administrativas do DF: o escore médio das localizações de Brasília fica entre os percentis 0,67 e 0,79 nas quatro.

Isso não invalida o escore: é possível que o pedestre ordene da mesma forma, e essa é precisamente a pergunta que a linha de base humana responderá. O que a observação impõe é uma restrição de desenho: qualquer validação futura, contra crime registrado ou contra resposta humana, precisa controlar a Região Administrativa. Sem esse controle, um modelo que apenas reconheça renda pelo asfalto e pelo acabamento das fachadas produzirá correlação alta e será lido como percepção de segurança.

O Qwen3-VL 8B específico é o contraexemplo útil, e a razão para não descartá-lo por ter confiabilidade menor: nele o escore médio de Brasília cai ao percentil 0,57, as 50 melhores são lideradas por Ceilândia e Taguatinga, e a fração urbana mediana das 50 piores desce para 0,70. É a única rodada cujo mapa não é substituível por um mapa de renda, o que mostra que a confusão das demais não é inevitável.

Uma segunda cautela decorre da distribuição dos escores: nas rodadas confiáveis, dezenas de imagens vencem ou perdem todos os seus confrontos, e o desempate entre elas cabe à força do adversário sorteado, que é ruído de pareamento. Apenas a faixa do ranqueamento é interpretável, e nenhuma imagem deste conjunto deve ser citada como "a mais segura".

V. Conclusões

Este trabalho apresentou uma metodologia de análise estruturada para avaliar a percepção de segurança urbana no Distrito Federal a partir de imagens no nível da rua e MLLMs, com validação cruzada prevista contra um padrão de referência humano local. Além do fluxo de trabalho, a etapa reportada entrega a caracterização empírica de um desenho fatorial de seis rodadas, com três modelos sob duas formulações de *prompt*.

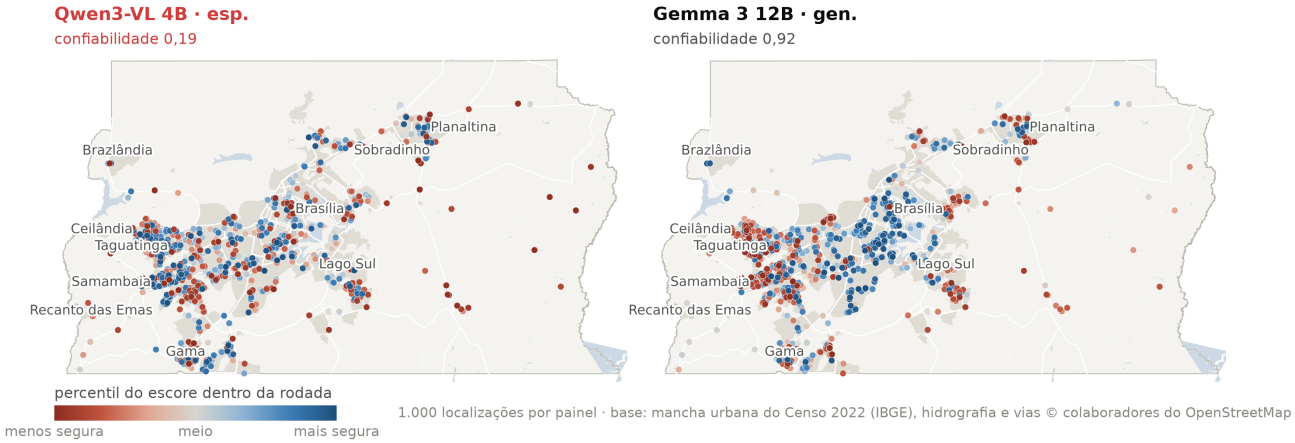


Fig. 4. O escore no espaço, na rodada de confiabilidade mais baixa e na mais alta. Cada ponto é uma das 1.000 localizações, colorido pelo percentil do seu escore dentro da própria rodada, já que as rodadas têm dispersões muito diferentes. À esquerda a cor não organiza o espaço; à direita o eixo Plano Piloto–Lago Sul aparece em azul e Ceilândia, Samambaia e Recanto das Emas em vermelho.

A confiabilidade do ranqueamento varia de 0,92 a 0,19 sob desenho idêntico e precisa ser estimada antes de qualquer correlação, por ser o seu teto. Ela cai sob a formulação mais longa nos dois modelos Qwen3-VL e em nenhum dos dois Gemma, padrão que aponta para a extração da resposta em modelos de raciocínio explícito como causa mais provável do que a dificuldade da pergunta. A formulação desloca 13% do sinal estável do Gemma 3 12B e quase 40% do Qwen3-VL 8B, o que a torna variável experimental, e modelos distintos sob a mesma pergunta compartilham cerca de metade da variância mesmo após correção pelo ruído, o que impede tratar "o MLLM" como entidade única.

A inspeção dos extremos revelou dois critérios visuais entre as rodadas confiáveis. Quatro delas ordenam por acabamento da via e da fachada e reproduzem a ordenação de renda entre Regiões Administrativas, restrição que qualquer validação futura precisará controlar; o Qwen3-VL 8B sob a formulação específica ordena por fachada ativa e movimento, e é a única cuja geografia não acompanha renda. O cruzamento com criminalidade registrada não sustentou associação após controlar escala e contaminação por geocodificação em centroide.

Como trabalhos futuros imediatos, destacam-se: (i) a conclusão da campanha de pesquisa de opinião e a medição da concordância humano-IA sob ambas as formulações, que decidirá qual dos dois critérios visuais se aproxima do pedestre real; (ii) a preservação dos votos brutos das rodadas de IA, com a posição de cada imagem no par, a resposta bruta e o motivo de encerramento, o que permitirá reestimar os escores sob Bradley-Terry, auditar viés posicional e testar a hipótese de truncagem; (iii) a extrapolação dos escores validados para toda a malha viária do DF por representações de visão-linguagem (CLIP/SigLIP) com K-NN; e (iv) o uso da base em busca em grafos para rotas seguras.

Referências

[1] J. Abraham *et al.*, "Crime and visually perceived safety of the built environment: A deep learning approach," *Annals of the American Association of Geographers*, 2025.

[2] H. Xie, L. Liu, H. Yue, "Modeling the effect of streetscape environment on crime using street view images and interpretable machine-learning technique," *International Journal of Environmental Research and Public Health*, vol. 19, no. 21, p. 13833, 2022.

[3] N. Naik *et al.*, "Streetscore: Predicting the perceived safety of one million streetscapes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014, pp. 779–785.

[4] J. Zhang, Y. Li, T. Fukuda, B. Wang, "Urban safety perception assessments via integrating multimodal large language models with street view images," *Cities*, 2025.

[5] O. Newman, *Defensible Space: Crime Prevention Through Urban Design*. New York: Macmillan, 1972.

[6] P. Salesses, K. Schechtner, C. A. Hidalgo, "The collaborative image of the city: Mapping the inequality of urban perception," *PLoS ONE*, vol. 8, no. 7, p. e68400, 2013.

[7] A. Dubey, N. Naik, D. Parikh, R. Raskar, C. A. Hidalgo, "Deep learning the city: Quantifying urban perception at a global scale," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 196–212.

[8] R. Herbrich, T. Minka, T. Graepel, "TrueSkill: A Bayesian skill rating system," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 19. MIT Press, 2007, pp. 569–576.

[9] R. A. Bradley M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[10] D. R. Hunter, "MM algorithms for generalized Bradley-Terry models," *The Annals of Statistics*, vol. 32, no. 1, pp. 384–406, 2004.

[11] L. Maystre M. Grossglauser, "Fast and accurate inference of Plackett-Luce models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 172–180.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[13] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, "Sigmoid loss for language image pre-training," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11 975–11 986.

[14] Y. Yan, H. Wen, S. Zhong, W. Chen, H. Chen, Q. Wen, R. Zimmermann, Y. Liang, "UrbanCLIP: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web," in *Proceedings of the ACM Web Conference (WWW)*, 2024.

[15] F. Moreno-Vera *et al.*, "Quantifying urban safety perception on street view images," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2021.

[16] A. Rzotkiewicz *et al.*, "Assessing google street view image availability in latin american cities," *Journal of Urban Health*, vol. 97, pp. 552–560, 2020.

[17] R. Mushkani, "Do vision-language models see urban scenes as people do? an urban perception benchmark," arXiv preprint arXiv:2509.14574, 2025.

[18] H. Lan, "Interpretable multimodal framework for human-centered street assessment: Integrating visual-language models for perceptual urban diagnostics," arXiv preprint arXiv:2506.05087, 2025.

[19] Z. Liu, T. Li, T. Ren, D. Chen, W. Li, W. Qiu, "Day-to-night street view image generation for 24-hour urban scene auditing using generative ai," *Journal of Imaging*, vol. 10, no. 5, p. 112, 2024.

[20] Gemma Team, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.

[21] Qwen Team, "Qwen3-VL," Alibaba Cloud. https://github.com/QwenLM/Qwen3-VL, 2025.