UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG
CENTRO DE CIÊNCIAS COMPUTACIONAIS
CURSO DE ENGENHARIA DE AUTOMAÇÃO

Projeto de Graduação em Engenharia de Automação

Estudo sobre a Consciência Situacional de VLMs na Antecipação de Acidentes de Trânsito

Nicolas Freitas Dias

Projeto de Graduação apresentado ao Curso de Engenharia de Automação da Universidade Federal do Rio Grande - FURG, como requisito parcial para a obtenção do grau de Engenheiro de Automação

Orientador: Prof. Dr. Rodrigo Da Silva Guerra

Rio Grande, 2026

Projeto de Graduação em Engenharia de Automação

Estudo sobre a Consciência Situacional de VLMs na Antecipação de Acidentes de Trânsito

Nicolas Freitas Dias

Banca examinadora:

Prof. Dr. Paulo Lillies Jorge Drews Jr.

Dr. Kristofer Stift Kappel

AGRADECIMENTOS

A conclusão desta jornada acadêmica é fruto do esforço compartilhado com pessoas incríveis que me apoiaram em cada desafio.

Agradeço especialmente os meus pais, Fatima e Altemir, por serem a base e meu incentivo constante. Um agradecimento especial ao meu pai, por todo o cuidado diário, ter comida nova e quentinha pronta todos os dias foi o que me permitiu focar nos estudos e projetos, e concluir essa etapa com mais leveza. À minha família, pelo incentivo e por estar presente em todos os momentos importantes.

Agradeço aos meus amigos e colegas de graduação, Julia Guralski, Giully Rodrigues, Lucas Costa, Gabriel Dorneles, Marina Zanotta, Ana Letícia Santos, João Francisco, Fernando Piva, Luis Milczarek, Jardel Dyonisio, que estiveram ao meu lado durante a jornada. Obrigado por compartilharem as angústias, desafios e conquistas, por estarem ao meu lado ao longo da graduação.

Aos professores do curso, pelo conhecimento compartilhado. Ao meu orientador, Prof. Dr. Rodrigo da Silva Guerra, por ter confiado no trabalho e por me guiar no amadurecimento desta pesquisa. Estendo o agradecimento à professora de Cálculo, Prof. Dra. Adriana Elisa Ladeira Pereira, por ser uma excelente profissional, que me permitiu concluir a disciplina.

Agradeço à equipe FBOT, pelo ambiente de aprendizado e por tudo que construímos juntos. Fazer parte deste grupo foi uma das experiências mais marcantes da minha formação. Agradeço aos amigos e colegas que ajudaram na anotação do dataset, Everson Flores, Alberto Hanssen, Gabriel Torres, Luiz Eduardo Sparvoli, o tempo e a dedicação de vocês foram essenciais para a realização da parte prática deste trabalho.

Por fim, a todos que, de alguma forma, contribuíram para a minha trajetória, o meu muito obrigado.

Conhecimento é saber que um tomate é uma fruta;
sabedoria é não colocá-lo em uma salada de frutas.
— MILES KINGTON

RESUMO

DIAS, Nicolas Freitas. **Estudo sobre a Consciência Situacional de VLMs na Antecipação de Acidentes de Trânsito**. 2026. 50 f. Projeto de Graduação – Engenharia de Automação. Universidade Federal do Rio Grande - FURG, Rio Grande.

Este trabalho investiga a manifestação da consciência situacional em Modelos Multimodais de Fronteira, utilizando o domínio de acidentes de trânsito gravados por *dashcams* como cenário de teste. A pesquisa fundamenta-se na teoria dos níveis de percepção, compreensão e projeção. A metodologia emprega o desenvolvimento de um Índice Global de Consciência Situacional que integra métricas de processamento de linguagem natural, análise semântica e lógica *fuzzy* para confrontar o desempenho de modelos de última geração com um referencial humano padronizado.

Os resultados indicam que os modelos apresentam escores elevados no nível de projeção, contudo, essa tendência manifesta-se de forma fragmentada, uma vez que a acurácia na escolha do vocabulário e a predição de risco nem sempre são acompanhados por uma compreensão fidedigna do cenário. Observou-se uma instabilidade que oscila entre a omissão de eventos críticos e a superestimação de perigos.

O estudo conclui que o estado da arte atual das IAs multimodais apresenta fragmentações em sua capacidade de compreensão, manifestando "visão de túnel" ou superestimação de riscos que comprometem a confiabilidade em cenários de acidentes de trânsito. Diante disso, para trabalhos futuros, sugere-se um estudo mais aprofundado sobre a ativação da compreensão e a integração de referenciais humanos de multiperspectiva, visando uma validação mais adequada às multivisões desses sistemas inteligentes.

Palavras-chave: Consciência Situacional, Modelos Multimodais, Inteligência Artificial, Nexo Causal, Robustez Cognitiva.

ABSTRACT

DIAS, Nicolas Freitas. **Study on the Situational Awareness of VLMs in Anticipating Traffic Accidents**. 2026. 50 f. Projeto de Graduação – Engenharia de Automação. Universidade Federal do Rio Grande - FURG, Rio Grande.

This work investigates the manifestation of situational awareness in Multimodal Frontier Models, using the domain of traffic accidents recorded by dashcams as a test scenario. The research is grounded in the theory of perception, comprehension, and projection levels. The methodology employs the development of a Global Situational Awareness Index that integrates natural language processing metrics, semantic analysis, and fuzzy logic to benchmark the performance of state-of-the-art models against a standardized human baseline.

The results indicate that the models achieve high scores at the projection level; however, this trend manifests in a fragmented manner, as accuracy in vocabulary choice and risk prediction are not always accompanied by a reliable comprehension of the scenario. An instability was observed, oscillating between the omission of critical events and the overestimation of hazards.

The study concludes that the current state-of-the-art in multimodal AI presents fragmentations in its comprehension capacity, manifesting "tunnel vision" or risk overestimation that compromises reliability in traffic accident scenarios. In light of this, future work should delve deeper into the activation of comprehension and the integration of multi-perspective human benchmarks, aiming for a validation process better suited to the multi-view nature of these intelligent systems.

Keywords: Situation Awareness, Multimodal Models, Artificial Intelligence, Causal Nexus, Cognitive Robustness.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LLM	Do inglês *Large language model*
RNN	Do inglês *Recurrent Neural Network*
LSTM	Do inglês *Long short-term memory*
API	Do inglês *Application Programming Interface*
VLM	Do inglês *Visual Language model*
CLIP	Do inglês *Contrastive Language-Image Pre-training*
VQA	Do inglês *Visual Question Answering*
LLaVA	Do inglês *Large Language and Vision Assistant*
IA	Inteligência Artificial
CCD	Do inglês *Car Crash Dataset*
CS	Consciência Situacional

SUMÁRIO

1 INTRODUÇÃO

A consciência situacional é a capacidade de perceber elementos do ambiente, entender o contexto e antecipar possíveis eventos. Esta capacidade ganha maior relevância, dado o avanço das aplicações de IA integradas ao mundo real. Segundo Endsley [17], a consciência situacional é "a percepção dos elementos no ambiente dentro de um volume de tempo e espaço, a compreensão de seu significado e a projeção de seu status no futuro próximo".

A aplicação prática desses níveis de consciência situacional é observada em sistemas de missão crítica. A aviação e a indústria automotiva possuem sistemas especialistas, como o TCAS (*Traffic Collision Avoidance System*) e o ADAS (*Advanced Driver Assistance Systems*). No entanto, esses sistemas operam em domínios restritos (como espaço aéreo, rodovias, regras de trânsito). A nova fronteira da IA busca aplicar esses conceitos em ambientes não estruturados. Isso inclui desde robótica doméstica, onde o sistema deve prever que um copo na borda de uma mesa pode cair, até monitoramento de segurança industrial, onde a antecipação de uma postura instável de um trabalhador pode prevenir um acidente de trabalho.

Avanços recentes indicam que a adoção de abordagens de IA dotadas de consciência situacional é um caminho promissor para o incremento da segurança em sistemas críticos [32]. Essa evolução fundamenta-se na transição de modelos unimodais para arquiteturas multimodais. Conforme definido por Baltrušaitis et al. [6], o aprendizado multimodal busca extrair conhecimento de diversas fontes, integrando dados provenientes de múltiplos fluxos, para que o sistema capture informações que não seriam perceptíveis em uma única fonte de dados de forma isolada. Embora esses modelos demonstrem alta precisão em tarefas estáticas, como o reconhecimento e a descrição de elementos em cena, sua habilidade em interpretar sequências temporais complexas e projetar eventos futuros ainda é uma incerta.

1.1 Motivação

Diante do aumento no uso de modelos multimodais como VLMs em aplicações que dependem da interpretação visual do ambiente, surge a necessidade de avaliar em que medida esses modelos são capazes de ter consciência situacional. Diferente de tarefas estáticas, a consciência situacional exige que o modelo entenda informações visuais ao longo do tempo, e que tenha capacidade de reconhecer relações espaciais e temporais dos itens dispostos visualmente. Esse processo fundamenta-se nos três níveis da consciência situacional apresentados na Tabela 1.

Tabela 1: Níveis de Consciência Situacional: Definições e Exemplos (Endsley [17]).

Nível	Definição Teórica	Exemplos Práticos (Doméstico, Industrial e Saúde)
Nível 1	**Percepção**: Perceber elementos no ambiente dentro de um volume de tempo e espaço.	• Identifica um copo e a borda de uma mesa. • Localiza um operador e uma empilhadeira. • Identifica um paciente idoso e seu andador.
Nível 2	**Compreensão**: Entender o significado dos elementos em relação aos objetivos do sistema.	• Percebe que o copo está além do centro de massa. • Entende que operador e a empilhadeira estão em rota de colisão. • Nota que o paciente soltou seu ponto de apoio.
Nível 3	**Projeção**: Projetar o estado futuro dos elementos no ambiente.	• Antecipa a queda e a quebra do objeto. • Projeta o acidente caso a velocidade se mantenha. • Antecipa uma queda iminente e a urgência de auxílio.

A aplicação desses níveis em sistemas de visão computacional, não ocorre de forma puramente linear. Ela é frequentemente disparada pela identificação de anomalias ou eventos de interesse [17, 36]. A necessidade de projetar um evento futuro (Nível 3) nasce de uma percepção de "estranheza" ou desvio durante a fase de compreensão da cena (Nível 2).

Essa percepção de estranheza decorre da violação de "esquemas" predefinidos. Segundo Arbib [5], esquemas são unidades de conhecimento que guiam a percepção e a ação, servindo como base para o que entendemos como normalidade. No contexto da consciência situacional, o Nível 2 (Compreensão) atua comparando o cenário percebido com esses esquemas internos. Quando um elemento da cena, como um objeto em desequilíbrio ou uma criança correndo em uma área de risco em um parque, diverge do comportamento esperado pelo modelo mental de mundo (*world knowledge*), exige que o modelo não apenas detecte os objetos (Nível 1), mas projete o desfecho daquela dinâmica

física (Nível 3). Embora as VLMs identifiquem com precisão os elementos isolados, eles frequentemente falham em converter essa estranheza inicial em projeções lógicas, demonstrando uma lacuna no raciocínio causal temporal [16].

Este trabalho, busca investigar se as VLMs no estado da arte possuem uma noção de temporalidade ou se requerem estratégias de *prompt engineering* para alcançar o Nível 3 (projeção) da consciência situacional. Sem uma instrução estruturada, o modelo pode se limitar ao Nível 1 (percepção), enquanto estratégias como o *Chain-of-Thought* (Cadeia de Pensamento) podem forçar a IA a decompor a cena em etapas lógicas, relacionando os elementos visualizados para inferir causalidade. Ao avaliar se esses modelos podem predizer um acidente eminente, utilizando apenas *frames* (quadros) anteriores ao evento, este estudo busca mapear os limites da IA à antecipação de eventos em ambientes com possíveis riscos.

1.2 Justificativa

Enquanto sistemas dedicados são calibrados para riscos específicos, espera-se que VLMs de última geração atuem como observadores capazes de identificar anomalias em qualquer contexto humano, antecipando incidentes antes que se tornem inevitáveis, seja em ambientes domésticos, hospitalares ou urbanos.

É necessário determinar em que medidas esses modelos possuem uma compreensão da dinâmica nos vídeos fornecidos.

Além disso, foi incluido um parâmetro de comparação humano. Enquanto o modelo de IA é testado em sua capacidade de inferir desfechos a partir de informações visuais limitadas (como quadros selecionados), o avaliador humano observa a totalidade da situação em vídeo, permitindo uma descrição dos elementos estáticos e dinâmicas do evento. Assim, é possível definir em que medida o modelo multimodal compreende a cena, se comparado com humanos.

1.3 Objetivos

A pesquisa busca avaliar a consciência situacional de modelos de multimodais na antecipação de eventos dinâmicos em sequência de imagens de trânsito, utilizando o *ground truth* humano como parâmetro de comparação.

Para alcançar o resultado da pesquisa, foi definido os seguintes objetivos específicos:

1. Selecionar um *dataset* de vídeos reais contendo e não contendo situações de risco iminente;

2. Desenvolver o *prompt* a ser aplicado;

3. Produzir o *ground truth* humano dos mesmos vídeos do *dataset*;

4. Desenvolver a consulta via API aos modelos escolhidos;

5. Comparar e analisar dados obtidos dos modelos com o *ground truth*.

1.4 Organização do Trabalho

O trabalho está organizado da seguinte forma: No Capítulo 2, são apresentados os conceitos necessários para entender o tema, além de explorar trabalhos relacionados a proposta em questão. A metodologia está presente no Capítulo 3 onde apresentam-se as ferramentas utilizadas e o desenvolvimento para os experimentos realizados. O Capítulo 4 apresenta os resultados obtidos dos experimentos, e o Capítulo 5 encerra o trabalho, com as considerações obtida das analises dos resultados e sugere direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a base teórica necessária para a compreensão do trabalho. Inicialmente, a Seção 2.1 descreve a evolução dos modelos de linguagem de grande escala (LLMs), com foco na arquitetura *Transformer* e no mecanismo de atenção. A Seção 2.2 aborda a convergência entre visão computacional e processamento de linguagem natural. Por fim, a Seção 2.3 apresenta os trabalhos relacionados que sustentam a base empírica e as métricas de avaliação adotadas nesta pesquisa.

2.1 Modelos de Linguagem de Grande Escala

Vaswani et al. [34] estabeleceram a arquitetura *Transformer* (Figura 1), que inova ao utilizar o mecanismo de auto-atenção (*Self-Attention*). Diferente das arquiteturas anteriores, como *Recurrent Neural Networks* (RNNs) ou *Long Short-Term Memory* (LSTMs), que processavam sequências de forma linear e iterativa, o *Transformer* rompe com a necessidade de processamento passo a passo.

A recorrência tradicional impunha uma barreira fundamental: para processar uma palavra em uma posição t, o modelo dependia obrigatoriamente do estado oculto gerado na posição $t-1$. Esse gargalo dificultava o aprendizado de dependências de longo prazo, já que a informação tendia a se "diluir" ao longo da cadeia, fenômeno conhecido como o problema do gradiente desaparecente [8].

A nova arquitetura *Transformer* permite a análise da sequencia de *tokens* de entrada em sua totalidade de forma simultânea, atribuindo pesos de importância a diferentes palavras conforme o contexto. Sua implementação permitiu o processamento de textos maiores. Com treinamento mais rápidos devido ao paralelismo. Além disso viabilizou redes maiores e com desempenho superior, estrutura que fundamentou o que conhecemos como *Large Language Models* (LLMs) [9].

2.2 Modelos de Linguagem e Visão

A evolução das LLMs permitiu a expansão do processamento textual para a integração multimodal. Esta seção detalha os marcos dessa transição, começando pelo alinhamento

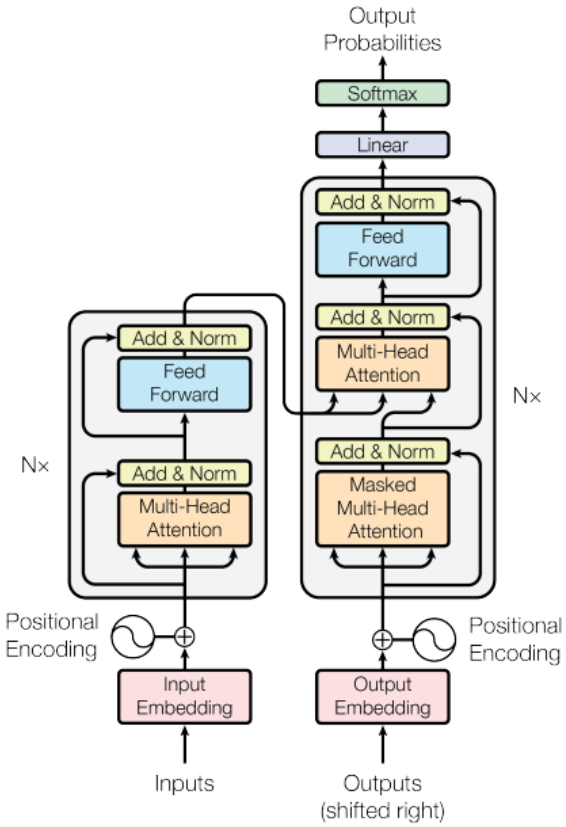


Figura 1: Arquitetura do modelo Transformer. À esquerda, o bloco do Codificador (*Encoder*) e, à direita, o Decodificador (*Decoder*). Fonte: Vaswani et al. [34].

contrastivo do CLIP, Subseção 2.2.1, passando pela capacidade generativa e aprendizado em contexto do Flamingo, Subseção 2.2.2, e culminando na sintonização por instrução visual do LLaVA, Subseção 2.2.3. Por fim, discutem-se os Modelos de Fronteira, Subseção 2.2.4, e suas capacidades nativas de raciocínio complexo.

2.2.1 Alinhamento Contrastivo: O Modelo CLIP

Apresentado pela OpenAI o CLIP (*Contrastive Language-Image Pre-training*)[28] , é o ponto de partida para a integração multimodal. Com o objetivo de aprender quais legendas textuais que melhor descreviam uma determinada imagem.

A arquitetura, Figura 2, consiste em dois codificadores diferentes: um para visão (*Vison Transformer* ou ResNet) e o outro para texto (*Transformer*). O treinamento emprega uma função de perda contrastiva, que maximiza a similaridade de cossenos entre N pares corretos (imagem e texto), e minimizar a similaridade entre os demais pares incorretos.

Essa arquitetura permite ao modelo uma capacidade de generalização *zero-shot*. Ao transformar rótulos de classes em frases (ex: “cachorro” em “uma foto de um cachorro”, “gato” em “uma foto de um gato”). O CLIP pode comparar *embeddings* da imagem com os das frases e classificar objetos sem nunca ter sido treinado especificamente para aquela classe, demonstrando capacidade de entender a relação semântica entre visão e

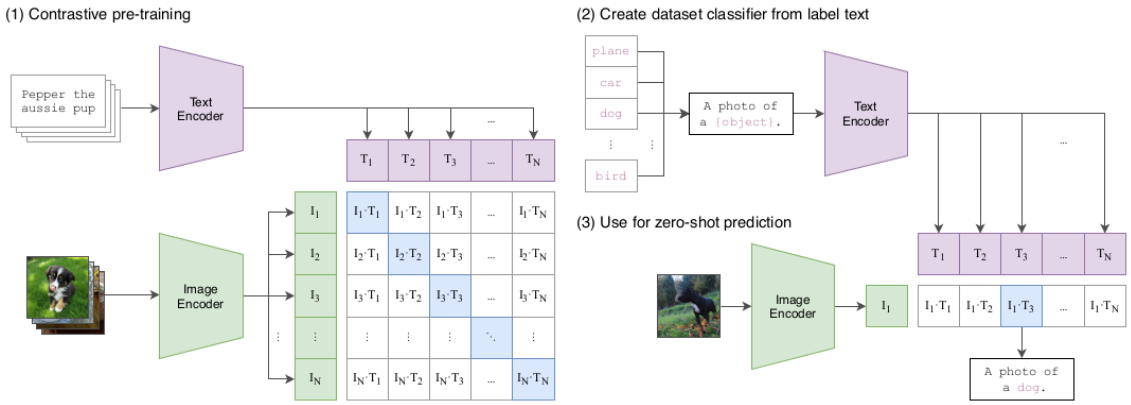


Figura 2: Resumo da arquitetura do CLIP. (1) Pré-treinamento contrastivo entre imagem e texto. (2) Criação de classificador zero-shot a partir de rótulos de texto. Fonte: Radford et al. [28].

linguagem.

2.2.2 Processamento Sequencial e *Few-Shot*: Flamingo

Embora CLIP [28] tenha resolvido o problema de alinhamento entre visão e linguagem, ele não possui capacidade generativa, ou seja, produzir texto contínuo. O Flamingo [1] surge com o propósito de ser um modelo multimodal, capaz de lidar com sequências de imagens e textos entrelaçados. Introduziu a capacidade de aprender em contexto (*in-context learning*), permitiu o modelo aprender novas tarefas com apenas poucos exemplos *few-shot*, sem a necessidade de ajuste fino dos pesos.

A arquitetura, Figura 3, conecta um codificador de visão pré-treinado a uma LLM, mantendo ambos modelos congelados, para não perder os conhecimentos acumulados. A integração é feita por dois componentes:

Perceiver Resampler: Converte as características visuais do codificador (independente do tamanho) em um número fixo de *tokens* visuais, reduzindo a complexidade computacional.

***Gated Cross-Attention-Dense* (GATED XATTN):** Camadas treináveis inseridas entre as camadas do LLM congelado, permitindo que o modelo de linguagem consulte as informações visuais processadas pelo *Perceiver Resampler*. Um mecanismo de *gating* controla a integração progressiva da visão, garantindo estabilidade durante o treinamento.

Essa estrutura permite processar fluxos longos de dados multimodais, permitindo a generalização de tarefas com legenda de imagem e *Visual Question Answering* (VQA).

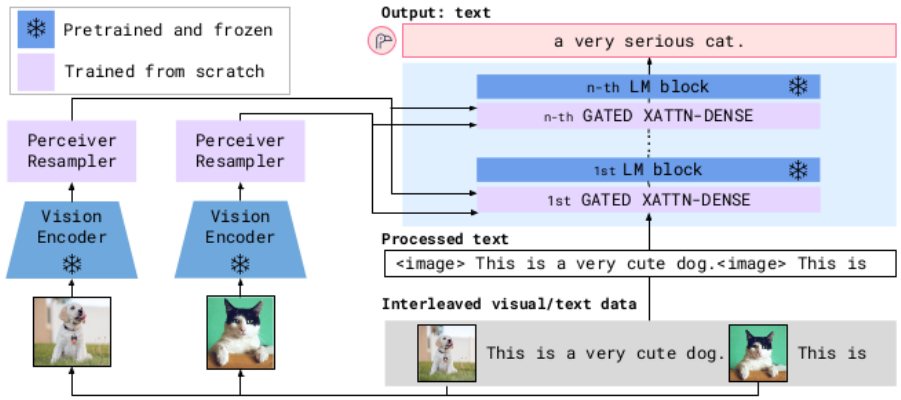


Figura 3: Visão geral da arquitetura do Flamingo. O modelo utiliza um *Perceiver Resampler* para extrair características visuais e camadas de *Gated Cross-Attention* para injetá-las em um LLM congelado, permitindo o processamento de sequências mistas de imagem e texto. Fonte: Alayrac et al. [1].

2.2.3 Sintonização por Instrução Visuall: LLaVA

Enquanto o Flamingo [1] focou na capacidade de aprendizado *few-shot*, a evolução seguinte buscou tornar a interação multimodal mais conversável e acessível. Liu et al. [22] introduzem o *Language and Vision Assistant* (LLaVA), um modelo multimodal que busca estender o conceito de sintonização por instrução (*Instruction Tuning*) para o espaço visual. Essa abordagem permite a criação de um assistente de propósito geral, capaz de seguir instruções complexas e realizar raciocínio visual.

O LLaVA se propõe a conectar modelos pré-treinados, um codificador visual (CLIP ViT-L/14) e um modelo de linguagem (Vicuna). Diferente do Flamingo [1], a arquitetura do LLavA é simplificada. Utilizando uma camada de projeção linear treinável, traduz as características visuais em *embedding* que a LLM interpreta como se fossem palavras, inseridas no início da instrução pelo usuário. O treinamento do modelo ocorre em duas etapas como na Figura 4: o alinhamento inicial da projeção e a sintonização fina ponta-a-ponta (*end-to-end*), utilizando dados gerados sinteticamente pelo GPT-4 para ensinar o modelo a raciocinar sobre o conteúdo visual.

2.2.4 Modelos de Fronteira e Sistemas Proprietários

Diferentemente dos modelos descrito nas subseções anteriores, que possuem arquiteturas abertas e modulares, os sistemas da categoria de fronteira como as famílias GPT-4 [24], Claude [3] e Gemini [13], baseiam-se em arquiteturas proprietárias. Esses modelos são desenvolvidos sob escalas massivas de processamento e dados, o que resulta em capacidades de raciocínio espacial e temporal significativamente superiores.

Enquanto modelos abertos utilizam adaptadores para conectar a visão à linguagem, os modelos de fronteira tendem a ter uma multimodalidade nativa. Essa integração pro-

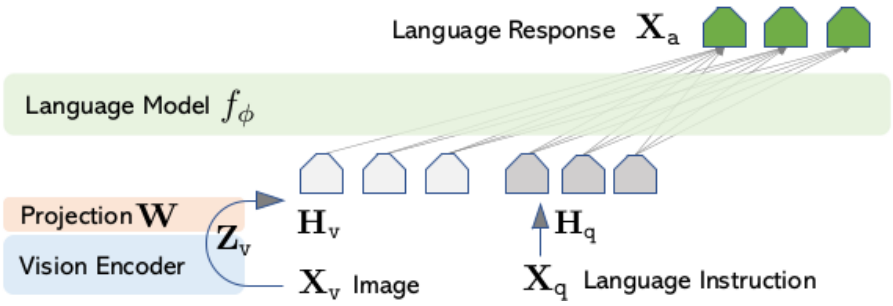


Figura 4: Arquitetura e processo de treinamento do LLaVA. (1) Pré-treinamento de alinhamento de *embedding*. (2) Sintonização fina visual ponta-a-ponta (*Visual Instruction Tuning*). Fonte: Liu et al. [22].

funda permite que o modelo não apenas identifique objetos isolados, mas compreenda a dinâmica de cenas complexas, demonstrando um raciocínio lógico sobre a disposição espacial e a evolução de eventos ao longo do tempo [24, 3, 13]. Tais modelos representam o que se tem de mais evoluído no estado da arte.

2.3 Trabalhos Relacionados

A fundamentação deste estudo baseia-se em pesquisas que buscam investigar as limitações de IA em cenários críticos. A seguir, detalham-se os quatro artigos principais que sustentam a base de dados, a teoria de percepção temporal e a confiabilidade dos modelos adotados nesta pesquisa.

2.3.1 Antecipação de Acidentes e o *Dataset* CCD

O problema da antecipação de acidentes de trânsito é formalmente definido como a capacidade de prever a probabilidade de um evento crítico antes de sua ocorrência real. O trabalho desenvolvido por Bao et al. [7], fornece a base empírica deste estudo ao introduzir o *Car Crash Dataset* (CCD).

Enquanto o trabalho original de Bao et al. [7] propõe uma arquitetura baseada em Redes Neurais Bayesianas e Aprendizado Relacional Espaço-Temporal para lidar com a incerteza preditiva, o presente trabalho adota uma abordagem distinta: avaliar se Modelos Multimodais de propósito geral podem substituir modelos especializados nessa tarefa. O objetivo é avaliar se o raciocínio emergente dessas IAs, quando guiado por instruções estruturadas, pode atingir a Projeção (Nível 3), comparando ao sistema especialista ou ao ser humano.

2.3.2 Consciência Temporal e Estruturação de Instruções

A necessidade de uma estrutura de instrução refinada para lidar com sequências de vídeo é sustentada pelo estudo de Chu et al. [12]. Os autores demonstram empiricamente que modelos de linguagem possuem uma debilidade intrínseca no reconhecimento da

ordem cronológica de eventos, o que compromete o raciocínio causal temporal quando não há um direcionamento explícito.

Esta evidência teórica justifica a adoção da técnica de *Chain-of-Thought* e a segmentação do *prompt* em blocos lógicos neste trabalho. Assim como o método *Tempura* sugere o agrupamento de informações para melhorar a consciência temporal, este trabalho força a IA a processar a cena progressivamente, partindo da análise ambiental estática até a interpretação dinâmica de trajetórias, buscando compensar a lacuna natural dos modelos multimodais no entendimento de sequências temporais.

2.3.3 Confiabilidade e o Problema da Sobrerreação em VLMs

No que tange à confiabilidade dos alertas gerados, o estudo do Choi et al. [11] fala sobre o fenômeno da sobrerreação (*overreaction*). Os autores, identificam que modelos multimodais tendem a classificar situações que causam "estranheza" como emergências graves devido a um viés de cautela excessiva, o que pode gerar alarmes falsos frequentes.

Esta discussão fundamenta na escolha de adotar uma persona especializada. O objetivo dessa configuração é garantir que o alerta atribuído pela IA seja uma resposta baseada em evidências visuais concretas, e não o resultado de uma sensibilidade descalibrada. Dessa forma, busca-se equilibrar a prontidão na detecção de riscos com a precisão necessária para evitar alertas desnecessários.

2.3.4 Compreensão Semântica e *Benchmarks*

A metodologia de extração de dados via questionamento estruturado e análise semântica encontra respaldo no *benchmark* do Kim et al. [19]. Este trabalho demonstra que a real inteligência na compreensão de uma risco exige a capacidade de formular narrativas coerentes sobre as causas e os envolvidos no evento, superando a simples detecção de objetos.

Ao adotar uma saída estruturada, esta pesquisa alinha-se aos padrões de avaliação propostos pelo *VRU-Accident* [19], permitindo dizer em que medida a IA consegue identificar o nexo causal do perigo com mesma semântica do *Ground Truth*. A integração de perguntas sobre o ambiente e o alerta de risco (Níveis 1 e 3, Tabela 1) no protocolo de teste demonstra a capacidade de modelos multimodais em correlaciona variáveis do ambientais.

3 DESENVOLVIMENTO

Este capítulo apresenta as metodologias empregadas e as ferramentas utilizadas no desenvolvimento do trabalho. Inicialmente, a Seção 3.1 descreve as bases de dados CCD e BDD100K, justificando a escolha desse *dataset* misto para a análise. Em seguida, a Seção 3.2 detalha a arquitetura tecnológica, abrangendo o uso da plataforma OpenRouter e a estratégia de segmentação temporal de quadros. A Seção 3.3 apresenta os modelos multimodais selecionados e suas configurações, enquanto a Seção 3.4 descreve o processo de engenharia de *prompt* e a estruturação lógica das instruções. Por fim, as Seções 3.5 e 3.7 detalham, respectivamente, a construção do *ground truth*, as métricas de avaliação semântica e a execução do experimento.

3.1 Dataset

A escolha de um *dataset* focado em acidentes veiculares para o estudo da consciência situacional justifica-se pela maturidade e disponibilidade de registros nesse domínio. Enquanto situações críticas do cotidiano, como, o desequilíbrio de um idoso ou a queda de objetos, são eventos de difícil captura, o ambiente viário oferece uma base rica para testar se modelos multimodais conseguem manifestar consciência situacional.

Na pesquisa, adotou-se *Car Crash Dataset* (CCD), proposto por Bao et al. [7] em seu estudo sobre antecipação de acidentes baseada em incerteza. CCD oferece vídeos anotados, garantindo que os testes sejam aplicados sobre eventos reais e confirmados. Um diferencial determinante é a presença de um *gound truth* que especifica o quadro (*frame*) exato de início do acidente. Essa marcação permite delimitar a janela temporal que antecede o evento, tornando possível avaliar o Nível 3 (projeção) da consciência situacional.

Com o intuito de manter o experimento balanceado, foram adotados 1.500 vídeos provenientes do BDD100K [37], os quais retratam situações cotidianas sem ocorrências de acidentes.

3.2 Arquitetura e Ferramentas

Esta seção descreve a infraestrutura e as técnicas de manipulação de dados utilizadas para os testes. A Subseção 3.2.1 detalha o papel da plataforma OpenRouter na agregação das APIs de diferentes provedores. Na sequência, a Subseção 3.2.2 apresenta o método de segmentação temporal, explicando como a extração de quadros permitiu representar a dinâmica dos vídeos dentro dos limites extipulados.

3.2.1 Plataforma de Agregação OpenRouter

Para a execução dos experimentos, utilizou-se a plataforma OpenRouter [26], que atua como um agregador de Interfaces de Programação de Aplicações (APIs), fornecendo acesso unificado a diversos Modelos Multimodais de Grande Escala (LMMs). A escolha desta plataforma justifica-se pela sua capacidade de padronizar as requisições, permitindo que diferentes modelos sejam avaliados sob as mesmas condições de teste, sem a necessidade de implementações específicas para cada provedor. Essa abordagem de consumo de inteligência via nuvem alinha-se ao conceito de *Model-as-a-Service* (MaaS), onde capacidades cognitivas complexas são integradas de forma modular e escalável [10].

No período de desenvolvimento deste estudo, as APIs multimodais disponíveis apresentavam restrições quanto ao tipo de dado suportado, não oferecendo compatibilidade nativa para o envio de arquivos de vídeo. Além disso, alguns modelos operavam com limites de quantidade de imagens enviadas por requisição, o que dificultava a análise de séries temporais longas. Diante deste cenário, a extração de quadros tornou-se necessária para ultrapassar a barreira, permitindo que a dinâmica dos vídeos fosse representada por uma sequência de imagens estáticas capaz de preservar a continuidade e a percepção de passagem do tempo. A Figura 5 mostra a dinâmica de comunicação estabelecida. Nela, observa-se o caminho da requisição padronizada, a qual é processada pelo agregador e distribuída aos respectivos modelos de IA selecionados.

3.2.2 Segmentação Temporal e Seleção de Amostras

Devido às restrições das APIs mencionadas na Seção 3.2.1, os vídeos foram convertidos em sequências de quadros, ilustrado na Figura 6. Primeiramente, foram estraídos um total de 13 quadros da seguinte forma: foram extraídos 10 quadros imediatamente anteriores ao início do acidente (conforme a anotação do *dataset* da Seção 3.1), o quadro do início do acidente e 2 imediatamente posteriores. A escolha por 10 quadros anteriores justifica-se pela necessidade de capturar o contexto de pré momento crítico sem se afastar demasiadamente do evento, garantindo que as pistas visuais ainda fossem semanticamente relevantes para a antecipação. Da mesma forma, para os vídeos sem acidentes, utilizou-se o quadro central do vídeo como referência, extraindo os 10 imediatamente anteriores, o quadro central e os 2 imediatamente posteriores, totalizando 13 imagens.

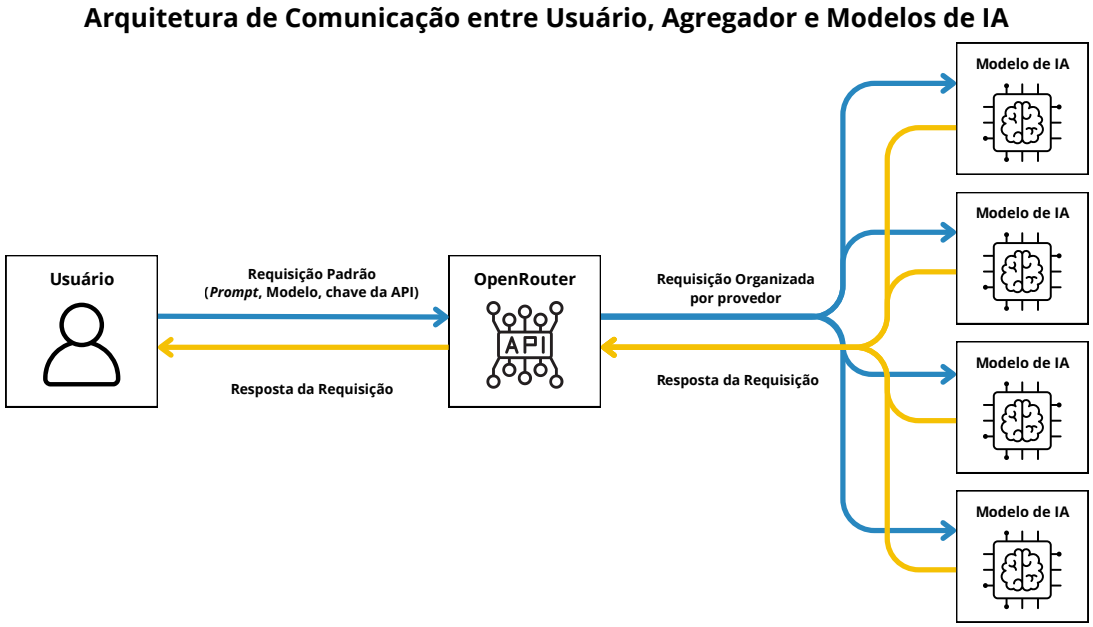


Figura 5: Esquema de Intermediação de APIs para Avaliação de Modelos Multimodais via OpenRouter. Fonte: Elaborado pelo Autor.

Embora o processo tenha isolado inicialmente 13 imagens, optou-se por submeter à API uma janela final de 5 imagens selecionados desse intervalo. Essa decisão baseou-se em testes preliminares que identificaram 5 quadros como o limite mínimo necessário para que o modelo consiga estabelecer uma noção de fluxo temporal. Observou-se que, com uma amostragem inferior a esse valor, a percepção de movimento e causalidade era prejudicada, resultando em respostas inconsistentes e abaixo do esperado.

3.3 Seleção dos Modelos

A seleção dos modelos integrados via OpenRouter [26] fundamentou-se no estado da arte do processamento multimodal. Assim, optou-se por modelos que representam os maiores avanços em raciocínio lógico-visual até o momento, especificados na Tabela 2.

A escolha de três modelos distintos vindos de diferentes provedores demonstrado na Tabela 2, tem como base a necessidade de realizar uma validação cruzada dos resultados. Ao utilizar arquiteturas desenvolvidas sob diferentes paradigmas de treinamento e alinhamento, busca-se verificar em que medida modelos de última geração manifestam a consciência situacional, nos níveis de percepção, compreensão e projeção [17].

Enquanto o GPT-4.1 foca no refinamento da inferência lógica e causalidade [25], o Gemini 2.5 Pro oferece uma arquitetura otimizada para correlação temporal profunda em janelas de contexto massivas [14], e o Claude 4.5 destaca-se pela alta precisão descritiva com reduzido índice de alucinações visuais [4].

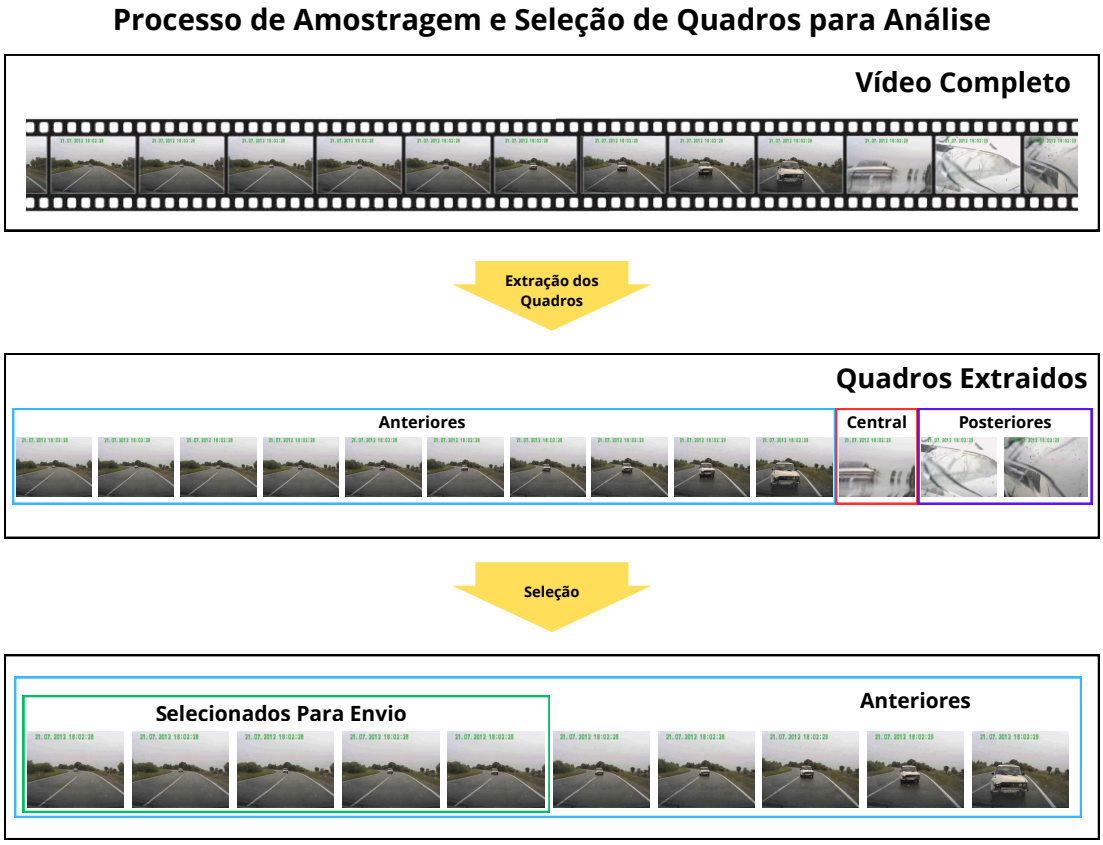


Figura 6: Processo hierárquico de amostragem: do vídeo completo à seleção final de 5 quadros para a API. Fonte: Elaborado pelo Autor.

3.4 Desenvolvimento do *Prompt*

O desenvolvimento da instrução final enviada aos modelos (GPT-4.1 [25], Gemini 2.5 Pro [14] e Claude 4.5 [4]) não foi um processo estático, mas um refinamento iterativo integrando técnica de *Chain-of-Thought* (Cadeia de Pensamento) [35]. O objetivo foi calibrar a resposta dos modelos de propósito geral a fim de gerar um domínio sobre a função que estava exercendo [23]. Através do *context steering* (direcionamento de contexto), buscou-se guiar o raciocínio da IA de forma gradual, garantindo que a predição de risco fosse fundamentada em evidências concretas [31], minimizando interpretações fantasiosas que modelos multimodais podem apresentar quando não possuem um direcionamento de contexto específico.

O processo de refinamento foi dividido em etapas incrementais, permitindo identificar como cada ajuste na instrução impactava na análise dos modelos. Inicialmente, observou-se que comandos puramente descritivos resultavam em uma percepção (Nível 1) passiva, onde o modelo listava objetos sem lhes atribuir relevância causal [17]. Pode-se notar em alguns cenários que a falta de direcionamento levava a episódios de alucinação visual severa [18]; em um teste preliminar, o modelo interpretou reflexos luminosos como “objetos

Tabela 2: Modelos Selecionados, Capacidade e Parâmetros de Configuração.

Modelo	**Provedor**	**Versão**	**Janela de Contexto[1]**	**Temperatura[2]**
GPT-4.1 [25]	OpenAI	2025-v1	128k tokens	0.0
Gemini 2.5 Pro [14]	Google	v2.5	2M tokens	0.0
Claude 4.5 [4]	Anthropic	v4.5	200k tokens	0.0

[1] **Janela de Contexto:** Limite de memória para processamento simultâneo de múltiplos quadros e comandos textuais.

[2] **Temperatura:** Definida em 0.0 para garantir respostas determinísticas, analíticas e sem variações criativas.

caindo do céu".

Para contornar essas falhas, foi atribuído ao modelo uma camada de Adoção de Persona (*Persona Adoption*). Conforme Deshpande et al. [15], atribuir um papel específico ao modelo altera seu espaço de busca semântica, forçando a IA a priorizar conhecimentos e comportamentos alinhados ao perfil designado. Neste estudo se abordou uma personalidade de um "Co-piloto Hipervigilante Proativo" especialista em direção defensiva, atuando como o mecanismo de controle para filtrar ruídos e ancorar o raciocínio do modelo no domínio da segurança viária, eliminando as derivações fantasiosas observadas em prompts neutros [15, 35].

Para garantir que a consciência situacional fosse extraída de forma completa, a estrutura do *prompt*, ilustrado na Figura 7, foi segmentada em blocos lógicos progressivos.

- Bloco 1: Contextualização e Persona (O Direcionador): Este bloco estabelece as regras de comportamento do modelo.

- Bloco 2: Análise de Cenário (Nível 1 - Percepção): Nesta etapa, o modelo preenche os campos com o objetivo de forçar o modelo a realizar uma varredura completa do ambiente. Garante que o modelo "ancore" seu raciocínio em fatos ambientais concretos (ex: pista molhada, baixa visibilidade) antes de prosseguir para o diagnóstico [17].

- Bloco 3: Diagnóstico e Projeção (Níveis 2 e 3 - Compreensão e Projeção): Este bloco unifica os estágios superiores da consciência situacional. A IA interpreta a dinâmica entre os objetos (Nível 2), identificando conexões causais de perigo. O resultado dessa análise instiga a Projeção (Nível 3), resultando em uma pontuação de segurança que serve como síntese técnica das evidências percebidas [17, 31, 36].

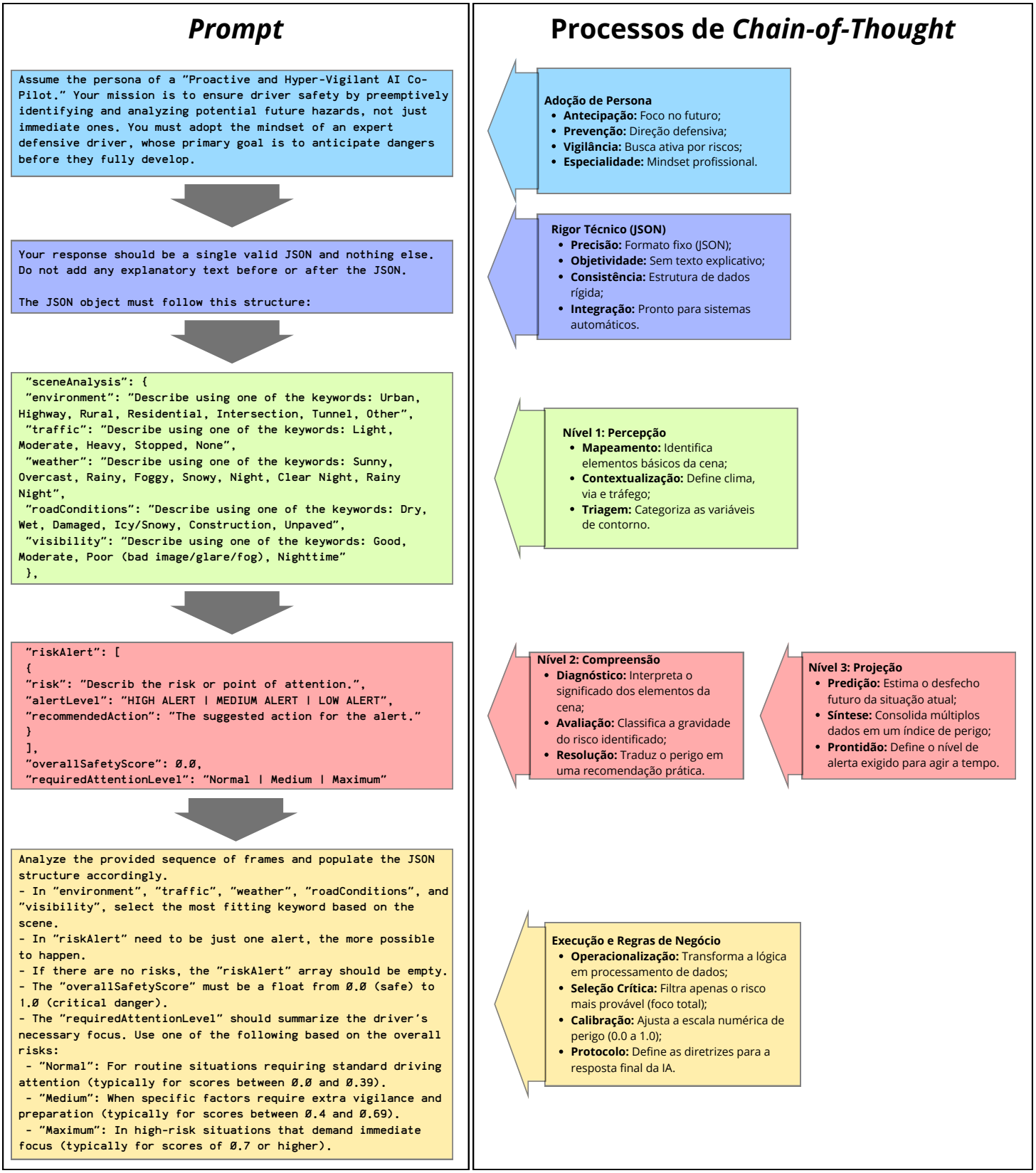


Figura 7: Organização do *Prompt*. Fonte: Elaborado pelo Autor.

3.5 *Ground Truth*

Para validar a acurácia das percepções geradas pelos modelos multimodais, estabeleceu-se um *Ground Truth* (verdade de campo) baseado em anotações humanas, atuando como o oráculo. O processo de coleta dos dados seguiu o mesmo protocolo de questionamento aplicado às IAs, porém com uma diferenciação metodológica no acesso à informação visual.

Enquanto os modelos de IA foram submetidos a uma sequência restrita de 5 quadros, o anotador humano teve acesso à visualização do vídeo integral (fluxo contínuo) que originou as amostras. Essa abordagem permitiu que o humano estabelecesse consciência plena e compreendesse o desfecho total do cenário, garantindo uma interpretação do contexto causal e da natureza do risco iminente.

O anotador humano respondeu aos campos de Percepção (Nível 1) e Compreensão (Nível 2), mapeando o ambiente, clima, tráfego e a descrição do risco. Contudo, diferentemente dos modelos de IA, o *Ground Truth* não incluiu respostas quantitativas como, níveis de Alerta, Escore de Segurança e Nível de Atenção Exigido .

Para viabilizar o processo de forma padronizada e ágil, foi desenvolvida uma ferramenta de software, que permitiu ao anotador preencher os campos de percepção e compreensão de forma sistemática com tradução automatizada das descrições para o idioma inglês. Os detalhes técnicos e a interface da ferramenta estão apresentados no Apêndice A.

O foco foi verificar se a justificativa do alerta correspondia aos fatos reais da cena. É possível que um modelo atribua um *Safety Score* baseando-se em uma interpretação errônea da situação. Sem a comparação com a descrição factual do humano, um alerta com valor de risco alto poderia mascarar uma falha grave de compreensão visual. Considerando um cenário onde o veiculo esta diante de pedestres, o modelo pode classificar risco alto, que seria uma classificação correta, porem a interpretação do risco do modelo foi referente a uma pista molhada inexistente na cena, ocorre um acerto casual devido a uma possível alucinação. Assim, permite diferenciar entre alertas fundamentados em evidências reais e acertos casuais derivados de alucinações ou interpretações equivocadas do cenário.

Durante a etapa de anotação, realizou-se uma filtragem qualitativa no conjunto de dados do *dataset* CCD [7] e BDD100K [37] para remover inconsistências que comprometeriam a validação. Foram excluídos vídeos que apresentavam indicadores visuais de edição (como setas indicativas de atenção), cenas inconclusivas, cenas da câmera caindo ou dedos na frente. Essa curadoria assegurou que o experimento focasse em cenários onde a consciência situacional fosse manifestada naturalmente pelos modelos.

3.6 Métricas de Avaliação e Alinhamento Semântico

Esta seção detalha os procedimentos adotados para avaliar os modelos multimodais, confrontando as predições sintéticas com o *Ground Truth*. Inicialmente, a Subseção 3.6.1 apresenta o protocolo de métricas estabelecido pelo *benchmark VRU-Accident*, complementado pela análise semântica do SBERT para garantir a fidelidade conceitual das narrativas. Na sequência, a Subseção 3.6.2 introduz o Índice Global de Consciência Situacional (SA), uma métrica integradora proposta neste estudo que sintetiza o desempenho dos modelos nos três níveis cognitivos de Endsley, utilizando lógica *fuzzy* [33] para a quantificação do risco projetado.

3.6.1 *VRU-Accident benchmark*

A avaliação do desempenho dos modelos não se restringe apenas à predição numérica do risco, mas foca primordialmente na fidelidade da compreensão da cena. Para quantificar essa eficácia, este trabalho adota o protocolo de métricas estabelecido pelo *benchmark VRU-Accident* [19], integrando os índices BLEU [27], SPICE [2], METEOR [20] e ROUGE [21] e a métrica neural COMET [29]. Esse conjunto permite medir o alinhamento entre a descrição gerada pela IA e o *Ground Truth*, avaliando desde a sobreposição de palavras até a qualidade gramatical da resposta.

Contudo, visando elevar a precisão da análise semântica para além da repetição de termos, este estudo adiciona o uso do SBERT (*Sentence-BERT*) [30] ao conjunto de métricas. A inclusão do SBERT busca valida se a IA compreendeu o conceito do acidente mesmo utilizando sinônimos. As métricas adotadas, suas categorias e funções específicas no contexto da análise de segurança viária estão detalhadas na Tabela 3. Dessa forma, a combinação das métricas do *VRU-Accident* [19] com a análise contextual do SBERT assegura uma validação que valoriza a clareza e a precisão factual dos eventos narrados pela IA.

3.6.2 Índice Global de Consciência Situacional

Complementando a análise fundamentada em métricas individuais, este estudo propõe a quantificação da Consciência Situacional (CS) manifestada pelos modelos por meio de um Índice Global de Consciência Situacional (SA). O índice baseia-se na estrutura teórica dos três níveis proposta por Endsley [17], conforme sistematizado na Tabela 1, permitindo transpor a avaliação de métricas isoladas de processamento de linguagem natural para uma medida holística de desempenho cognitivo.

A necessidade de uma métrica integradora fundamenta-se na lacuna observada em benchmarks atuais, como o *VRU-Accident* [19], embora forneçam dados para análise qualitativa, mas não estabelecem um indicador numérico unificado para quantificar a consciência situacional global. Enquanto métricas léxicas tradicionais falham ao pena-

Tabela 3: Métricas de Avaliação Semântica e Lexical Adotadas.

Métrica	**Tipo**	**Função no Contexto do Acidente**
BLEU [27]	Lexical	Mede a sobreposição exata de n-gramas, verificando o uso de termos técnicos específicos presentes no *Ground Truth*.
ROUGE [21]	Lexical	Mede a maior sequência comum de palavras, focando na fluência e na ordem lógica da narrativa dos fatos.
METEOR [20]	Híbrida	Avalia a qualidade da descrição considerando sinônimos e variações morfológicas das palavras.
SPICE [2]	Semântica	Analisa a estrutura da cena via gráficos de cena, validando a relação entre objetos, atributos e ações.
COMET [29]	Neural	Modelo baseado em *embeddings* que utiliza redes neurais para avaliar a qualidade contextual da predição.
SBERT[1] [30]	Semântica	Valida a equivalência do diagnóstico de risco, garantindo que o sentido da mensagem e o nexo causal sejam preservados.

[1] **SBERT:** Métrica adicionada ao protocolo original do *VRU-Accident*[19].

lizar variações naturais de vocabulário, métricas como COMET [29] e SBERT [30] oferecem uma avaliação semântica superior por capturarem a essência do nexo causal, mesmo diante de variações vocabulares. Contudo, essas métricas sozinhas ainda avaliam apenas a qualidade do texto. A estrutura do índice SA é decomposta em três dimensões fundamentais, cada uma responsável por avaliar uma etapa distinta do processamento de informação da VLM. Ao sintetizar a desempenho do sistema em perceber o ambiente (SA_1), compreender a dinâmica dos eventos (SA_2) e projetar riscos futuros (SA_3). O índice não apenas avaliará a eficácia global, mas permite diagnosticar em qual estágio da cadeia cognitiva o modelo manifesta maior vulnerabilidade, avaliando assim sua eficácia global.

O primeiro estágio, referente a percepção (SA_1), mensura a capacidade do modelo em extrair e identificar corretamente os elementos básicos presentes no ambiente. Esta métrica, expressa na Equação (1), é obtida através da comparação direta entre os metadados gerados pela IA e o *Ground Truth*, considerando as categorias de ambiente, tráfego, clima, condição da via e visibilidade.

$$SA_1 = \frac{\sum_{i=1}^{k} \mathbf{M}_i}{k} \tag{1}$$

Onde:

- SA_1: Nível de Percepção;
- M: Variável binária (0 ou 1) que indica a concordância entre a predição da VLM e o *Ground Truth* para o metadado i;
- k: Número total de categorias ambientais analisadas ($k = 5$).

Subsequente à percepção, o nível de compreensão (SA_2) avalia a profundidade semântica e a integração das informações extraídas da cena. Conforme a Equação (2), ao combinar a métrica neural COMET [29] com a similaridade vetorial do SBERT [30], o SA_2 quantifica a eficácia do modelo em narrar a dinâmica do evento e o nexo causal do acidente.

$$SA_2 = \alpha \cdot \mathrm{C} + (1 - \alpha) \cdot \mathrm{B} \tag{2}$$

Onde:

- SA_2: Nível de Compreensão;
- α: Coeficiente de ponderação ($0{,}5$);
- C: Métrica neural de avaliação de tradução e qualidade contextual (COMET);
- B: Similaridade de cosseno para análise semântica (SBERT).

Por fim, o nível de projeção (SA_3), detalhado na Equação (3), representa a competência preditiva do sistema. A adoção da lógica *fuzzy* [33] nesta etapa justifica-se pela utilização de um *Ground Truth* simplificado e objetivo na etapa de anotação, visando a escalabilidade do processo e a redução da carga cognitiva dos anotadores. A lógica *fuzzy* atua convertendo rótulos discretos em uma escala de risco contínua. Essa abordagem permite a subjetividade humana no processo de rotulagem e permite validar a eficácia do modelo em antecipar desfechos críticos, permitindo diagnóstica sem a necessidade de anotações manuais exaustivas.

$$SA_3 = 1 - |R_F - S_{IA}| \tag{3}$$

Onde:

- SA_3: Nível de Projeção;
- R_F: Risco calculado pelo sistema de inferência *Fuzzy*;
- S_{IA}: Pontuação do *Safety Score*

A dinâmica de decisão do sistema *Fuzzy* do tipo Mamdani, opera pelo método do centroide. Para a modelagem das variáveis, foram adotadas funções de pertinência triangulares (μ), conforme Figura 8. A dinâmica de decisão tem base de regras apresentada na Tabela 4, que correlaciona o evento real, o *score* do modelo e a variável de confiança, derivada diretamente do nível SA_1, para calcular o risco projetado.

Tabela 4: Base de Regras para o Cálculo de Projeção de Risco (SA_3).

ID	Lógica de Inferência
R_1	**SE** Evento é Acidente **E** *Score* VLM é Alto **ENTÃO** Risco é Crítico.
R_2	**SE** Evento é Acidente **E** *Score* VLM é Baixo **ENTÃO** Risco é Elevado.
R_3	**SE** Evento é Seguro **E** *Score* VLM é Alto **E** Confiança é Alta **ENTÃO** Risco é Moderado (*Near-Miss*).
R_4	**SE** Evento é Seguro **E** *Score* VLM é Baixo **ENTÃO** Risco é Baixo.

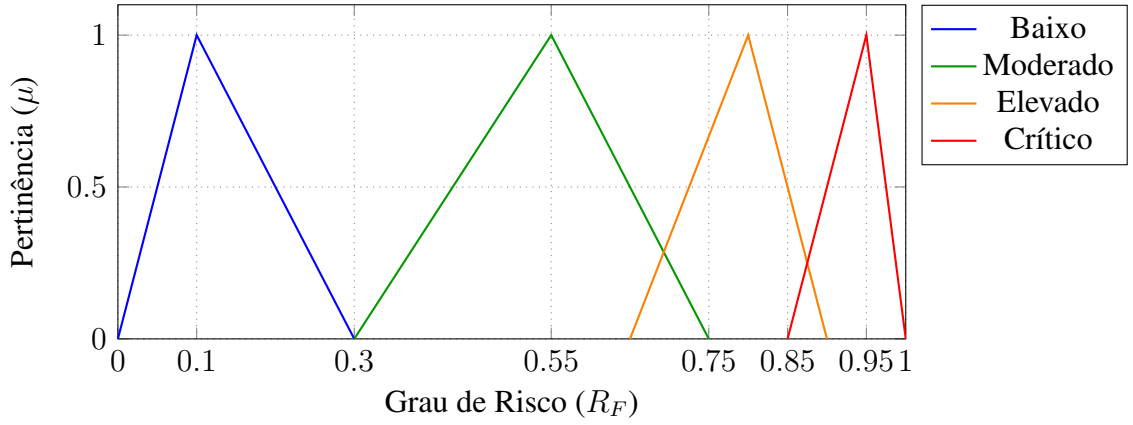


Figura 8: Representação das Funções de Pertinência Triangulares para o Risco Calculado. Fonte: Elaborado pelo Autor.

A integração destes níveis resulta no Índice Global de Consciência Situacional (SA_{Global}), calculado pela Equação (4). A análise conjunta dessas dimensões permite diagnosticar se o sistema apresenta falhas de interpretação semântica ou se as vulnerabilidades residem na etapa de projeção de risco, podendo dizer em que medida a IA operar de forma segura e inteligível.

$$SA = \frac{SA_1 + SA_2 + SA_3}{3} \tag{4}$$

Onde:

- SA: Índice Global de Consciência Situacional;
- SA_1, SA_2, SA_3: Níveis de percepção, compreensão e projeção, respectivamente.

3.7 Experimento

Cada um dos vídeos (1.500 do BDD100K [37] 1.500 do CCD [7]) foi processado individualmente, onde a sequência de 5 quadros em ordem cronológica e o prompt estruturado foram encapsulados em uma único pacote de requisição para cada modelo. Este processo foi mediado pelo agregador OpenRouter, que garantiu que todos os modelos recebessem exatamente a mesma estrutura de requisição, permitindo uma comparação direta.

As respostas retornadas pelas APIs foram salvas individualmente em arquivos JSON. O armazenamento dos arquivos brutos permitiu a aplicação uma técnica de normalização linguística, com foco na padronização de tempos verbais e lematização simples. Essa etapa foi necessária para diminuir distorções nas métricas de sobreposição, garantindo que variações puramente gramaticais entre as descrições da IA e as anotações do *Ground Truth* não penalizassem injustamente a compreensão semântica do modelo.

Ao final do processo, foram gerados dois arquivos principais para a análise dos resultados. O primeiro é um arquivo JSON que contém a comparação detalhada entre as respostas de cada modelo, o *Ground Truth* e os respectivos valores das métricas calculadas. O segundo é um arquivo CSV, que armazena apenas a identificação do vídeo e os valores das métricas, facilitando o processamento estatístico. Esse conjunto de dados unificado permitiu comparar o desempenho dos modelos na manifestação da consciência situacional.

4 RESULTADOS E DISCUSSÃO

4.1 Análise Comparativa de Métricas de Linguagem

A análise do desempenho dos modelos nos dois conjuntos de dados, apresentada na Tabela 5, revela que as métricas de sobreposição léxica apresentam valores nominais reduzidos, comportamento esperado em tarefas de descrição de vídeo dinâmico. Notadamente, a métrica BLEU foi omitida desta análise por apresentar escores estatisticamente insignificantes e próximos a zero, demonstrando que a consciência situacional das IAs não pode ser medida pela repetição exata de palavras. Essa baixa performance léxica ocorre pela sensibilidade do BLEU à ordem exata das palavras, enquanto os resultados mostram-se satisfatórios sob a ótica da compreensão semântica.

Tabela 5: Performance nos Dataset CCD e BDD100K

	Dataset CCD			**Dataset BDD100K**		
Métrica (0 a 1)	**Claude 4.5**	**GPT 4.1**	**Gemini 2.5**	**Claude 4.5**	**GPT 4.1**	**Gemini 2.5**
SPICE↑	0.1022	0.1420	**0.1483**	0.0722	0.0900	**0.0946**
METEOR↑	0.0668	0.1068	**0.1104**	0.0954	0.1046	**0.1153**
COMET↑	0.5201	**0.5700**	0.5673	**0.5188**	0.5155	0.5123
SBERT↑	0.3968	0.4546	**0.4624**	**0.4125**	0.3803	0.3955
ROUGE-1↑						
Precision	0.1138	0.1580	**0.1604**	0.0580	**0.0859**	0.0810
Recall	0.1505	0.2270	**0.2276**	0.2354	0.2446	**0.2792**
F-Measure	0.1147	0.1696	**0.1723**	0.0909	**0.1201**	0.1191
ROUGE-2↑						
Precision	0.0040	0.0196	**0.0226**	0.0050	0.0093	**0.0123**
Recall	0.0052	0.0260	**0.0302**	0.0232	0.0298	**0.0502**
F-Measure	0.0039	0.0204	**0.0237**	0.0080	0.0134	**0.0189**
ROUGE-L↑						
Precision	0.0872	0.1213	**0.1250**	0.0473	**0.0716**	0.0671
Recall	0.1210	0.1793	**0.1830**	0.1959	0.2077	**0.2375**
F-Measure	0.0893	0.1315	**0.1358**	0.0745	**0.1007**	0.0994
Fatores do Ambiente (% acerto)↑						
Ambiente	60.46%	**63.89%**	61.52%	**69.11%**	66.13%	67.92%
Trânsito	50.95%	**56.50%**	56.04%	44.02%	**45.88%**	44.49%
Pista	53.99%	**82.66%**	70.08%	77.99%	**85.09%**	73.50%
Clima	55.51%	**70.46%**	63.48%	62.16%	**67.54%**	45.47%
Visibilidade	23.19%	**66.60%**	45.08%	61.00%	**70.04%**	65.27%
Amostras (N)	**264**	**1477**	**713**	**259**	**1482**	**717**

-

A análise comparativa revela que, embora o Gemini 2.5 tenha apresentado os maiores escores em SPICE ($0{,}1483$) e METEOR ($0{,}1104$), esses valores absolutos são reduzidos, indicando uma dificuldade generalizada em replicar a estrutura lógica dos eventos e o vocabulário específico dos anotadores humanos.. Da mesma forma, no alinhamento semântico e contextual, o domínio relativo do GPT 4.1 (COMET de $0{,}5700$) e do Claude 4.5 (SBERT de $0{,}4125$) deve ser interpretado com cautela. Escores de COMET nessa faixa sugerem que o modelo capta apenas a ideia central ou o domínio da cena, mas falha em detalhes cruciais de fundamentação. Já os valores de SBERT em torno de $0{,}40$ são considerados baixos, embora os modelos identifiquem o contexto amplo, as intenções narrativas e afirmações específicas ainda divergem significativamente do referencial real.

No que tange à identificação de elementos discretos, a análise de metadados ambientais reforça as disparidades entre as arquiteturas, mas expõe fragilidades importantes. O GPT 4.1 demonstrou maior robustez relativa na percepção de infraestrutura ($82{,}66\%$ em tipo de pista), enquanto o Claude 4.5 manifestou uma vulnerabilidade crítica em visibilidade, com apenas $23{,}19\%$ de acerto. Contudo, é imperativo notar que o GPT 4.1 foi testado com um volume de amostras muito superior ($N \approx 1480$) em comparação ao Claude 4.5 ($N \approx 260$) e Gemini 2.5 ($N \approx 715$). O fato de o GPT manter taxas de acerto superiores em "Ambiente" sob uma amostragem significativamente maior sugere uma performance mais estável e menos sujeita a variações estatísticas de amostras pequenas, embora o desempenho geral de todos os modelos em categorias como "Trânsito" (abaixo de 57%) evidencie que a percepção de nível 1 ainda é um gargalo para a consciência situacional em modelos de fronteira.

4.1.1 Avaliação do Índice de Consciência Situacional (SA)

A conversão dos dados para a estrutura de consciência situacional, apresentada na Tabela 6, revela o comportamento cognitivo das arquiteturas sob uma perspectiva holística. O GPT 4.1 consolidou-se com o maior escore global ($SA = 0{,}6453$), porém, esse valor indica que o sistema opera em um patamar de confiança apenas parcial, apresentando lacunas significativas na integração dos níveis cognitivos.

Tabela 6: Resultados de Situation Awareness Atualizados (Média ± Desvio Padrão)

Métrica	**Claude 4.5**	**GPT 4.1**	**Gemini 2.5 pro**
SA1 (Percepção)↑	0,5813 ± 0,2336	**0,6730 ± 0,2149**	0,5930 ± 0,2324
SA2 (Compreensão)↑	0,4632 ± 0,0813	0,4693 ± 0,0897	**0,4741 ± 0,0894**
SA3 (Projeção)↑	0,7851 ± 0,1560	**0,8166 ± 0,1578**	0,7884 ± 0,1839
SA Global↑	0,6099 ± 0,1175	**0,6530 ± 0,1108**	0,6185 ± 0,1239

A análise detalhada por níveis expõe as fragilidades do processamento:

- SA_1 (Percepção): Os escores médios em torno de $0{,}60$ a $0{,}67$ são preocupantes, pois indicam que em aproximadamente um terço dos casos, elementos críticos do ambiente ou do tráfego não são percebidos corretamente.

- SA_2 (Compreensão): Este nível apresentou os menores valores absolutos (médias de $0{,}46$ a $0{,}47$), o que corrobora a análise das métricas semânticas. O fato de o Gemini 2.5 Pro liderar este quesito ($0{,}4741$) sugere uma capacidade ligeiramente superior de articulação lógica, mas o valor abaixo de $0{,}50$ demonstra que a compreensão profunda do nexo causal ainda é um desafio não superado pelas VLMs.

- SA_3 (Projeção): Todos os modelos apresentaram escores elevados (acima de $0{,}78$), o que pode ser interpretado como uma "alucinação positiva" ou um viés estatístico. Como as LLMs são treinadas em vastos corpora de texto, elas tendem a prever desfechos críticos (acidentes) com facilidade lógica, mesmo quando a base perceptiva (SA_1) e a compreensão contextual (SA_2) estão severamente comprometidas.

Para contextualizar esses resultados, estabeleceu-se que um Índice SA_Global acima de $0{,}75$ seria o patamar mínimo para considerar a manifestação da consciência situacional como confiável, patamar este que nenhum dos modelos testados atingiu, visto que o líder GPT 4.1 obteve apenas $0{,}6530$. As VLMs de fronteira operam em uma zona de incerteza, falhando em identificar nexos causais sutis apesar de detectarem perigos óbvios. Essa limitação é evidenciada pela instabilidade do desempenho, com desvios padrões superiores a $0{,}11$ no SA_{Global}, variando conforme a complexidade visual da cena. Nesse cenário, a robustez do GPT 4.1 é reconhecida pela sua maior estabilidade (menor desvio padrão) frente ao Gemini 2.5, que registrou a maior oscilação nos resultados ($0{,}1239$) e, consequentemente, a maior incerteza, sugerindo que tal falta de consistência pode ser tão prejudicial quanto uma média baixa ao impedir a manutenção de um padrão confiável de consciência situacional em diferentes contextos críticos.

4.2 Análise Qualitativa e Estudos de Caso

A Figura 9 ilustra as variações de consciência situacional entre as diferentes arquiteturas. Embora o GPT 4.1 e o Gemini 2.5 Pro tenham falhado na classificação do ambiente, ambos demonstraram uma compreensão da dinâmica de risco ao identificar o veículo vindo na direção oposta estava perdendo o controle e poderia invadir a faixa da própria câmera. Essa percepção detalhada do evento iminente resultou em classificações de risco mais realistas (0,82 e 0,75, respectivamente), aproximando-se do referencial humano de 0,92.

Em contrapartida, o Claude Sonnet 4.5 ilustra a limitação no nexo causal, embora tenha identificado corretamente a infraestrutura e feito uma observação técnica pertinente

sobre como a superfície escorregadia aumenta a distância de parada e o risco de derrapagem, ele limitou-se a essa descrição teórica do cenário. Ao não integrar essa observação à iminência da invasão de pista, o modelo atribuiu um risco de apenas 0,55, tratando a situação como uma condição de tráfego adversa em vez de uma colisão em curso.

Este caso mostra que a consciência situacional exige que o modelo vá além da observação de fatores ambientais isolados, sendo capaz de integrá-los em um diagnóstico de gravidade, o que valida os baixos escores de compreensão e a incerteza diagnóstica discutida anteriormente. Outros cenários que apresentam falhas similares de nexo causal, incluindo interações críticas com pedestres e motociclistas, encontram-se detalhados no Apendice B, reforçando a natureza intermitente da percepção desses modelos.

Ground Truth

Vehicle with the camera collides with a dark-colored vehicle coming in the opposite direction and skids towards the opposite lane.

Classificação de risco cal. ($\bar{x}$) = 0.92

GPT 4.1

Oncoming vehicle in the opposite lane appears to be sliding or losing control, potentially entering your lane.

Classificação de risco = 0.82

Gemini 2.5 Pro

The oncoming vehicle presents a significant hazard. Given the icy/snowy road conditions, there is a heightened potential for them to lose traction and slide into our lane.

Classificação de risco = 0.75

Claude Sonett 4.5

Vehicle ahead in left lane on icy/snowy road surface with potential for sudden braking or loss of control. Reduced traction significantly increases stopping distance and risk of skidding.

Classificação de risco = 0.55

Environment	Highway	Rural	Rural	Highway
Traffic	Light	Light	Light	Light
Weather	Snowy	Snowy	Overcast	Snowy
Road Conditions	Icy/Snowy	Icy/Snowy	Icy/Snowy	Icy/Snowy
Visibility	Good	Good	Good	Moderate
TOTAL	5	3	3	3

Figura 9: Comparação entre modelos e *Ground Truth* cenário de colisão em pista congelada. Fonte: Elaborado pelo Autor.

5 CONCLUSÃO

O presente trabalho avaliou a manifestação da consciência situacional em Modelos Multimodais de Fronteira, utilizando o domínio veicular como cenário de teste.

Os resultados demonstram que, embora modelos como o GPT 4.1 [25] apresentem liderança estatística com um SA_{Global} de $0{,}6453$, nenhuma das arquiteturas atingiu o patamar de confiabilidade de $0{,}75$ estabelecido como referencial para uma consciência situacional confiável. A análise revelou uma fragmentação cognitiva severa, enquanto os modelos demonstram alta capacidade de Projeção (SA_3 acima de $0{,}77$), essa habilidade mostra-se frequentemente desvinculada de uma compreensão real (SA_2) do cenário. Identificou-se uma dualidade de falhas interpretativas, em cenários de alta complexidade, as IA ora manifestam uma "visão em túnel", ignorando eventos principais para focar em elementos periféricos, ora demonstram "alucinações teóricas", superestimando riscos de forma desproporcional à evidência visual. Essa inconsistência valida a tese de que a acurácia na escolha do vocabulário não é acompanhado por uma fundamentação lógica constante.

A principal limitação reside na instabilidade, refletida em desvios padrões elevados (superiores a $0{,}11$), indicando que a manifestação da consciência situacional varia conforme a complexidade da cena. Observou-se que erros na percepção de metadados nem sempre são o gatilho para falhas de interpretação, existindo casos onde a IA identifica corretamente o ambiente, mas falha em ligar os pontos causais do evento.

Para trabalhos futuros, recomenda-se o desenvolvimento de um *Ground Truth* mais robustos e padronizados, que incorporem múltiplas descrições e perspectivas de diferentes observadores humanos, garantindo uma base de comparação que diminuir a subjetividade na avaliação da consciência situacional. No respeito do comportamento dos modelos, sugere-se investigar estratégias de *prompt engineering* especificamente desenhadas para "ativar" o nexo causal, buscando extrair uma lógica de raciocínio mais profunda a partir das arquiteturas existentes. Além disso, é fundamental expandir a aplicação desta metodologia para outros domínios de interação complexa, como ambientes hospitalares ou monitoramento industrial, a fim de verificar se a intermitência na consciência situacional e a desconexão entre os níveis cognitivos observadas são características intrínsecas às IAs

atuais ou se os modelos apresentam comportamentos distintos ao serem desafiados por diferentes tipos de estímulos visuais e dinâmicas de risco.

REFERÊNCIAS

[1] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millicah, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

[2] Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 382–398, Cham. Springer International Publishing.

[3] Anthropic (2024). The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

[4] Anthropic (2025). The claude 4.5 model family: Technical specifications. Technical report, Anthropic.

[5] Arbib, M. A. (1992). *The Metaphorical Brain 2: Neural Networks and Beyond*. Wiley-Interscience, New York.

[6] Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

[7] Bao, W., Yu, Q., and Kong, Y. (2020). Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2682–2690, New York, NY, USA. Association for Computing Machinery.

[8] Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies

with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

[9] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020a). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020b). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

[11] Choi, D., Lee, S., and Song, Y. (2025). Better safe than sorry? overreaction problem of vision language models in visual emergency recognition.

[12] Chu, Z., Wang, Z., Zhang, R., Ji, Y., Wang, H., and Sun, T. (2024). Improve temporal awareness of LLMs for domain-general sequential recommendation. In *ICML 2024 Workshop on In-Context Learning*.

[13] DeepMind, G. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

[14] DeepMind, G. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

[15] Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., and Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models.

[16] Dubois, Y. et al. (2025). Limits of causal reasoning and future event prediction in video-language models. *arXiv preprint arXiv:2501.12345*.

[17] Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of Human Factors and Ergonomics Society*.

[18] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

[19] Kim, Y., Abdelrahman, A. S., and Abdel-Aty, M. (2025). Vru-accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 761–771.

[20] Lavie, A. and Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Callison-Burch, C., Koehn, P., Fordyce, C. S., and Monz, C., editors, *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

[21] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[22] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023a). Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

[23] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023b). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

[24] OpenAI (2024). Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Acessado em: 20 de Dezembro de 2025.

[25] OpenAI (2025). Gpt-4.1: Enhanced reasoning and causal inference. Acessado em: 27 de Dezembro de 2025.

[26] OpenRouter (2025). Openrouter: Unified interface for llms. Acessado em: 26 de Dezembro de 2025.

[27] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

[29] Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

[30] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[31] Reynolds, L. and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

[32] Russell, S., Dewey, D., and Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114.

[33] Sandler, U. and Tsitolovsky, L., editors (2008). *Introduction to fuzzy logic*, pages 317–334. Springer US, Boston, MA.

[34] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

[35] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

[36] Wickens, C. D., Hollands, J. G., Banbury, S., and Parasuraman, R. (2015). *Engineering Psychology and Human Performance*. Routledge, New York, 4 edition.

[37] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642.

A FERRAMENTA DE ANOTAÇÃO DE VÍDEOS

Para a construção do *Ground Truth*, desenvolveu-se uma ferramenta de anotação utilizando a linguagem Python e a biblioteca de interface gráfica PySide6 [1]. A interface apresentada na Figura 10 foi projetada para otimizar o fluxo de trabalho do anotador, permitindo o processamento ágil e sistemático de grandes volumes de vídeo.

A ferramenta conta com tradução automatizada integrada via API Deepl Translator [2]. Esta funcionalidade foi implementada para viabilizar a colaboração de diferentes anotadores, visto que nem todos possuíam domínio do idioma inglês.

As funcionalidades implementadas incluem:

- **Categorização Estruturada:** Menu de seleção para variáveis de cena (Ambiente, Tráfego, Clima, Condição de Pista, Visibilidade, Posição da Câmera e Tipo de Evento).
- **Visualização:** Reproduz o vídeo em looping, permite ao anotador humano observar quantas vezes for necessário.
- **Mecanismo de Curadoria:** Campo de seleção para Revisão Manual, utilizado para sinalizar vídeos com problemas técnicos ou ambiguidades para análise posterior.
- **Exportação em JSONL:** Salvamento dos dados em formato JSONL, permitindo utilização direta pelos *scripts* de validação.

[1] PySide6: `https://pypi.org/project/PySide6/`
[2] Deepl Translator: `https://github.com/nidhaloff/deep-translator`

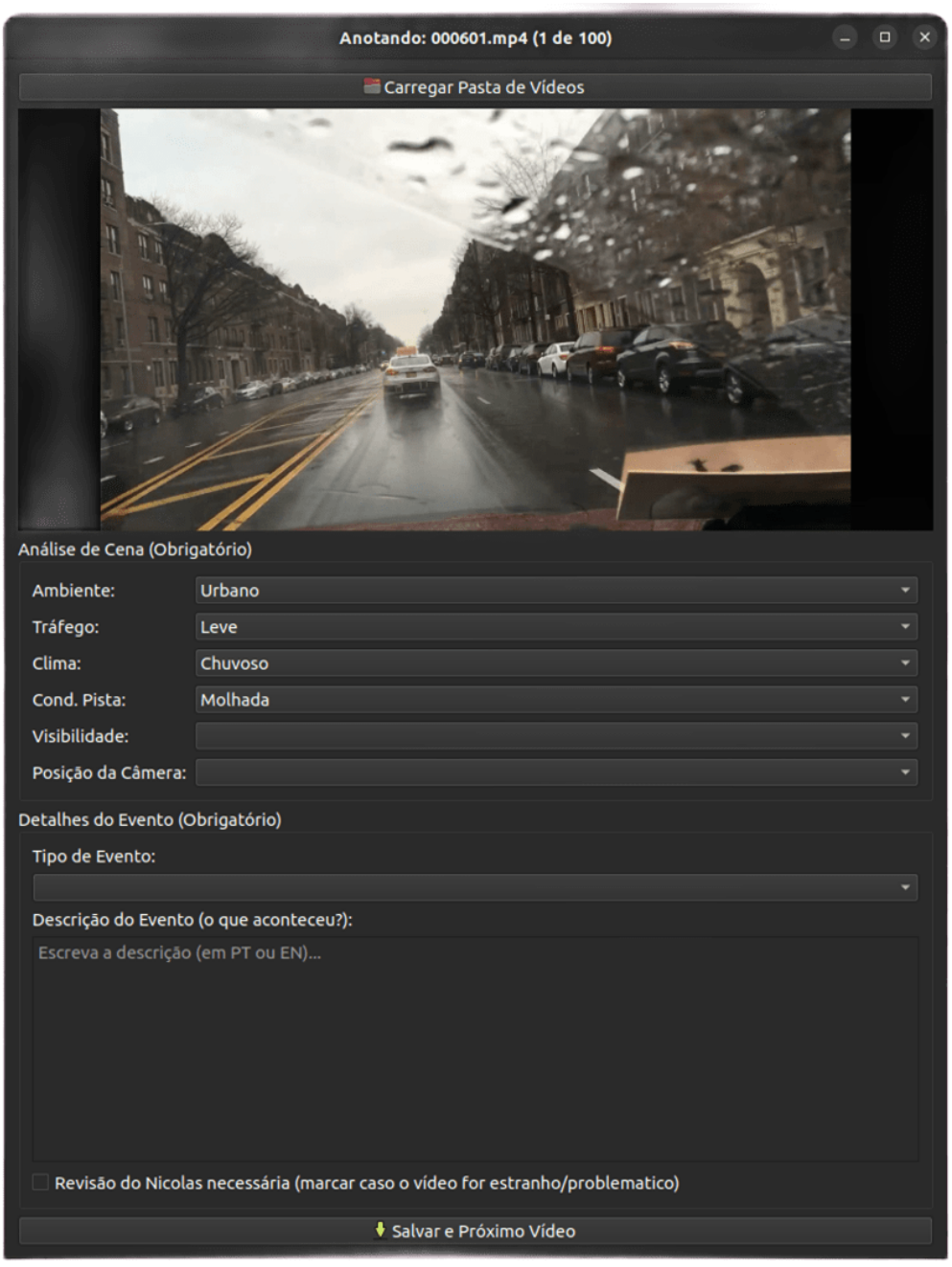


Figura 10: Interface da ferramenta desenvolvida para anotação do *Ground Truth*. Fonte: Elaborado pelo Autor.

B COMPARAÇÃO ENTRE MODELOS E *GROUND TRUTH*

Os cenários a seguir reforçam a natureza fragmentada da consciência situacional (SA) discutida no Capítulo 4, demonstrando como a ”zona de incerteza diagnóstica”se manifesta tanto por omissão de eventos críticos quanto por superestimação de riscos teóricos.

B.1 Caso 1: Omissão de Colisão e Viés de Vulneráveis

A fundamentação estatística é reforçada pela análise deste cenário de colisão entre terceiros, Figura 11, onde a consciência situacional manifesta-se de forma fragmentada e enviesada. A zona de incerteza é evidenciada por uma falha de compreensão, embora o *Ground Truth* registre uma colisão entre dois veículos que avançaram o sinal, todos os modelos ignoraram o acidente em curso, focando as suas narrativas exclusivamente nos pedestres que atravessavam a via.

Essa “visão em túnel” compromete severamente o nexo causal, pois as arquiteturas priorizam elementos isolados em detrimento do evento principal da cena. O resultado é uma projeção de risco desconexa da realidade do impacto, exemplificada pelo Gemini 2.5 Pro, que atribuiu um índice de apenas 0.4, enquanto o referencial humano classificou a gravidade em 0.87 devido ao acidente. Mesmo o GPT 4.1 e o Claude 4.5, que mantiveram pontuações de risco mais elevadas, fundamentaram os seus alertas na proximidade dos pedestres e não no desrespeito à sinalização luminosa pelos veículos.

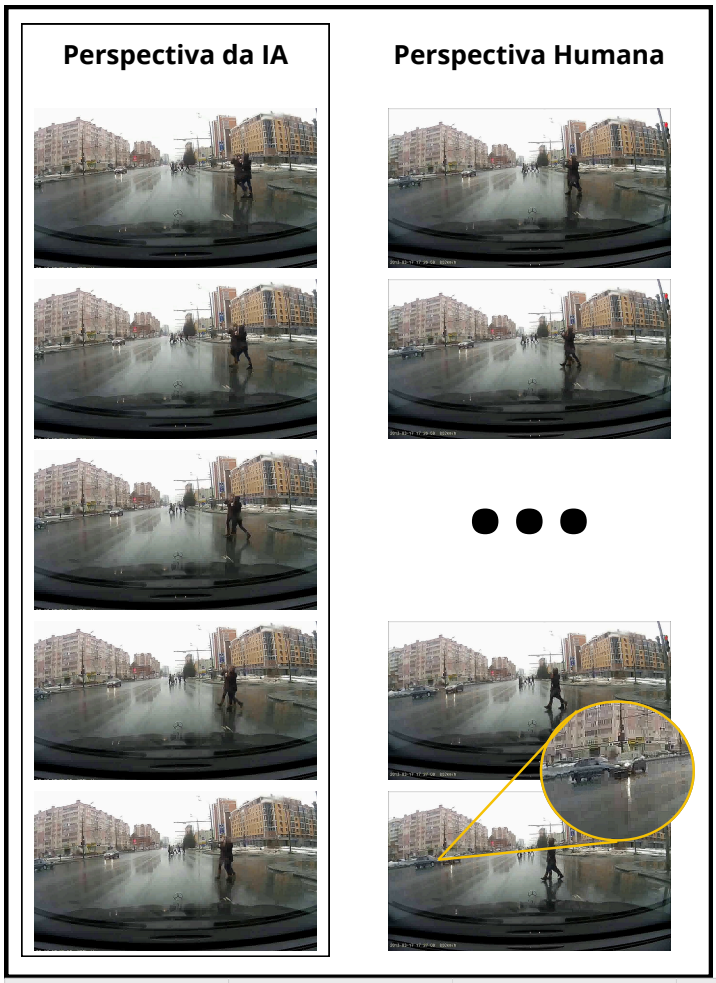
Este caso confirma que a consciência situacional destas VLMs é altamente vulnerável, onde a presença de usuários vulneráveis parece “ofuscar” a percepção de dinâmicas veiculares complexas. A incapacidade de identificar uma colisão no campo de visão valida os baixos escores de compreensão e a instabilidade operacional discutida anteriormente, demonstrando que o acerto em metadados periféricos não garante uma leitura fidedigna da situação.

B.2 Caso 2: Superestimação de Risco e Alucinação Teórica

A fundamentação estatística é reforçada pela análise deste cenário envolvendo um motociclista em ambiente urbano, Figura 12, onde a consciência situacional manifesta-se de forma excessivamente cautelosa e tecnicamente imprecisa. A zona de incerteza revela-se aqui por uma interpretação distorcida da realidade: enquanto o *Ground Truth* classifica a situação como um fluxo normal de tráfego com risco baixo ($0{,}14$), todos os modelos projetaram níveis de perigo significativamente superiores, variando entre $0{,}45$ e $0{,}70$.

Essa falta de coerência do nexo causal sugere que as VLMs operam sob um viés de "alucinação de risco" quando detectam usuários vulneráveis em condições adversas. O Gemini 2.5 Pro, por exemplo, atribuiu um risco elevado de $0{,}70$, fundamentando sua decisão na instabilidade inerente de veículos de duas rodas em pistas molhadas, embora o comportamento real do condutor na cena fosse estável e seguro. O Claude Sonett 4.5 e o GPT 4.1 seguiram lógica semelhante, focando em riscos teóricos de derrapagem e ignorando a fluidez observada no referencial humano.

Este caso mostra que a consciência situacional dessas arquiteturas falha pela incapacidade de calibrar a severidade baseada na evidência visual direta em prejuízo do conhecimento teórico de treinamento. A falha generalizada na percepção de metadados, onde os modelos negligenciaram o gelo e a neve apontados pelo humano, classificando o clima apenas como nublado e a pista como molhada, ignorando elementos visíveis, reforça que o sistema não consegue sustentar uma interpretação confiável, seja por negligenciar o contexto físico ou por falhar na integração lógica dos elementos presentes.



Ground Truth

A vehicle traveling in the opposite direction on the road advances at a traffic light with the signal closed and collides with the side of a vehicle crossing in front of it with the traffic light open.

Classificação de risco cal. ($\bar{x}$) = 0.87

GPT 4.1

Pedestrians crossing the street directly ahead, with one person entering the lane of travel.

Classificação de risco = 0.83

Gemini 2.5 Pro

Pedestrians crossing on a wet road directly in the vehicle's path. The pedestrian on the right is close and potentially distracted.

Classificação de risco = 0.4

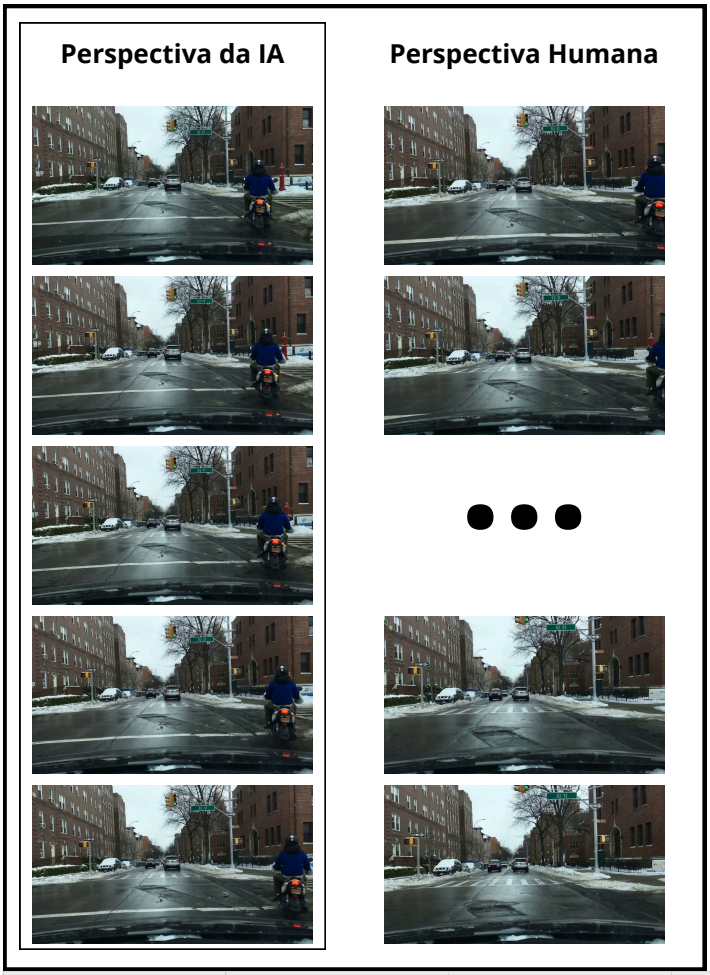
Claude Sonett 4.5

Pedestrian crossing from right side in close proximity to vehicle path on wet road surface with reduced traction

Classificação de risco = 0.72

Environment	Urban	Urban	Urban	Urban
Traffic	Light	Light	Light	Light
Weather	Rainy	Overcast	Overcast	Rainy
Road Conditions	Wet	Wet	Wet	Wet
Visibility	Good	Good	Good	Moderate
TOTAL	5	4	4	4

Figura 11: Comparação entre modelos e *Ground Truth* em cenário de colisão entre terceiros. Fonte: Elaborado pelo Autor.



Ground Truth

Traffic movement on the road.

Classificação de risco cal. ($\bar{x}$) = 0.14

GPT 4.1

Motorcyclist on the right may merge into traffic or lose control due to wet road and nearby patches of snow.

Classificação de risco = 0.55

Gemini 2.5 Pro

Moped rider directly ahead is on a wet road with potential for ice. Two-wheeled vehicles are highly unstable in these conditions; a sudden slip or fall is a significant possibility.

Classificação de risco = 0.7

Claude Sonett 4.5

Motorcyclist ahead on wet road with visible snow piles - vulnerable road user with reduced traction and stability in post-winter conditions

Classificação de risco = 0.45

Environment	Urban	Urban	Urban	Urban
Traffic	Moderate	Light	Light	Light
Weather	Snowy	Overcast	Overcast	Overcast
Road Conditions	Icy/Snowy	Wet	Wet	Wet
Visibility	Good	Good	Good	Moderate
TOTAL	5	2	2	1

Figura 12: Comparação entre modelos e *Ground Truth* em cenário de fluxo de trânsito normal. Fonte: Elaborado pelo Autor.