Avaliação da Capacidade de Detecção de Emoções em Modelos de Linguagem Multimodais (VLMs)

Gustavo Campos Pádua*
Prof. Dr. Ismael Santana Silva†

2026

Resumo

A capacidade de sistemas de Inteligência Artificial de compreender emoções é crucial para uma interação humano-computador empática e eficaz. Embora os Modelos de Linguagem Multimodais (VLMs) tenham demonstrado sucesso notável em tarefas objetivas, seu desempenho em tarefas subjetivas, como o reconhecimento de emoções, permanece pouco explorado. Esta pesquisa tem como objetivo avaliar e comparar o desempenho de VLMs de estado da arte no reconhecimento de emoções evocadas a partir de estímulos visuais. Adotando uma abordagem metodológica mista, utilizamos o *benchmark* EmoSet para avaliar quantitativamente as famílias de modelos Gemma 3 e Qwen3-VL sob execução local. Qualitativamente, analisamos os tipos de erros (e.g., viés de sentimento, excitabilidade incorreta) e o impacto de atributos visuais para compreender as limitações de raciocínio dos modelos. As contribuições esperadas incluem uma avaliação detalhada da inteligência emocional dos VLMs, o contraste em arquiteturas *zero-shot*, e *insights* para o comportamento de IAs generativas em processos de ajuste fino (*fine-tuning*) realizados em um modelo específico.

Palavras-chave: Inteligência Artificial. Visão Computacional. Modelos de Linguagem Multimodais. Reconhecimento de Emoções.

Abstract

The ability of Artificial Intelligence systems to understand emotions is crucial for empathetic and effective human-computer interaction. Although Multimodal Language Models (VLMs) have shown remarkable success in objective tasks, their performance in subjective tasks, such as emotion recognition, remains underexplored. This research aims to evaluate and compare the performance of state-of-the-art VLMs in recognizing

*gustavo.padua@aluno.cefetmg.br
†Orientador. Centro Federal de Educação Tecnológica de Minas Gerais – DECOM.

1 Referencial Teórico

Esta seção apresenta os conceitos principais utilizados para fundamentar a avaliação da capacidade de detecção de emoções por modelos de inteligência artificial. A seguir será apresentado um panorama sobre o reconhecimento de emoções na área de Visão Computacional, a complexidade focada no processamento de emoções evocadas, e, por fim, a arquitetura e funcionamento dos Modelos de Linguagem Multimodais (VLMs) que são objeto deste estudo.

1.1 Reconhecimento de Emoções em IA

O reconhecimento de emoções é uma área consolidada na Visão Computacional. As abordagens tradicionais dependem de processos rígidos de processamento visual para isolar características das imagens. Geralmente, o fluxo envolve a conversão do quadro capturado para escala de cinza, a localização geométrica utilizando classificadores computacionais baseados em recursos estruturais como o *Haar Cascade* e, por fim, o recorte e redimensionamento para dimensões padronizadas e compactas, frequentemente estabelecidas em 48×48 pixels (LIMA, 2023). A partir dessa segmentação dedicada, utilizam-se Redes Neurais Convolucionais (CNNs) para extrair características locais e mapeá-las em dimensões contínuas (valência e excitabilidade) ou categorias discretas de emoção (AKHAND et al., 2021; LIMA, 2023).

Modelos clássicos da psicologia estruturam essas classificações em conjuntos fixos de estados afetivos, variando desde as seis emoções básicas de Ekman (1993) até subdivisões contidas em bases consolidadas da área (BHATTACHARYYA; WANG, 2025). Enquanto esses métodos tradicionais focam estritamente em recortes isolados, os modelos baseados em arquiteturas multimodais contemporâneas conseguem assimilar o contexto global da cena sem a obrigatoriedade de etapas severas de segmentação prévia.

1.2 Modelos de Redes Convolucionais e Transfer Learning

Antes da consolidação dos modelos multimodais generativos de larga escala, o estado da arte em processamento visual afetivo baseava-se em técnicas de transferência de aprendizado (*transfer learning*) aplicadas a redes convolucionais profundas (LIMA, 2023). Arquiteturas amplamente consolidadas na literatura, como VGG16 e ResNet50 (HE et al., 2016), pré-treinadas em grandes bases de dados genéricas de classificação de imagens, como o ImageNet, eram utilizadas como robustos extratores de características visuais e representações semânticas fundamentais (LIMA, 2023).

Nesse paradigma convolucional clássico, as camadas iniciais aprendem padrões locais de baixa complexidade, como bordas e texturas faciais, enquanto camadas mais profundas capturam representações semânticas progressivamente mais abstratas. Nos próximos estágios, operações de subamostragem (*pooling*) reduzem a dimensionalidade espacial das ativações intermediárias, preservando informações discriminativas relevantes. Camadas adicionais adicionadas ao topo dessa base convolucional permitem adaptar a rede para tarefas específicas de reconhecimento visual e afetivo (LIMA, 2023).

1.3 Emoções Evocadas

No escopo desta pesquisa, é crucial focar no conceito de **Emoção Evocada**. Refere-se ao sentimento, reação ou humor que a totalidade da imagem desperta em seu observador

humano. Esta tarefa exibe enorme complexidade e subjetividade, visto que depende de fatores estéticos, semânticos, cromáticos e contextuais distribuídos por toda a imagem, extrapolando o domínio restrito de traços fisionômicos isolados (BHATTACHARYYA; WANG, 2025). Diferentemente de abordagens rasas que se limitam à identificação de um rosto sorrindo ou chorando na cena, a emoção evocada demanda uma compreensão holística e reflexiva de toda a composição fotográfica.

1.4 Discretização e Bases de Dados Afetivas

A avaliação empírica de arquiteturas de Aprendizado de Máquina no domínio afetivo impõe a padronização das classes de saída com base em conjuntos de dados anotados por múltiplos revisores humanos. Avaliações voltadas especificamente para o cenário de emoção evocada utilizam taxonomias expandidas, como a adotada no conjunto de dados de larga escala EmoSet, que categoriza os estímulos visuais em oito classes afetivas principais (diversão, admiração, contentamento, excitação, raiva, nojo, medo e tristeza), agregando também anotações detalhadas sobre atributos e contexto para sustentar o rigor da classificação subjetiva (YANG et al., 2023).

1.5 Modelos de Linguagem Multimodais (VLMs)

Para compreender o funcionamento dos Vision Language Models (VLMs), é necessário primeiramente definir as arquiteturas basilares que os compõem: os grandes modelos de linguagem e os codificadores visuais avançados. A base do processamento textual atual reside nos **Large Language Models (LLMs)**. Estes modelos são fundamentados na arquitetura *Transformer* (VASWANI et al., 2017), que revolucionou a inteligência artificial ao dispensar o uso de redes recorrentes e convolucionais puras em favor de mecanismos de autoatenção (*self-attention*). Essa arquitetura permite que o modelo identifique dependências globais e processe grandes sequências de texto paralelamente de forma altamente eficiente (VASWANI et al., 2017).

No domínio da Visão Computacional, essa capacidade de focar em partes específicas da informação foi adaptada através do **Vision Transformer (ViT)** (DOSOVITSKIY et al., 2021). O ViT fragmenta uma imagem bidimensional em uma sequência de quadros menores de tamanho fixo (*patches*, tipicamente de 16x16 pixels) e os processa linearmente, de forma análoga aos *tokens* (palavras) de um texto. Essa abordagem provou que a dependência estrita de Redes Convolucionais (CNNs) não é necessária, permitindo aplicar o modelo *Transformer* puro diretamente para o reconhecimento visual em larga escala (DOSOVITSKIY et al., 2021).

Para unificar as representações visuais e linguísticas, destaca-se a técnica de **Contrastive Language–Image Pretraining (CLIP)** (RADFORD et al., 2021). Treinado sobre uma base massiva de 400 milhões de pares de imagens e textos coletados da internet, o CLIP utiliza a aprendizagem contrastiva para treinar simultaneamente um codificador de imagem e um codificador de texto. O objetivo do treinamento é prever qual legenda textual corresponde a qual imagem, aproximando representações corretas e distanciando pares incorretos em um espaço matemático compartilhado (RADFORD et al., 2021).

Com o advento tecnológico dessas três bases, estabeleceram-se os Vision Language Models (VLMs). Esses modelos são arquiteturas multimodais que integram um codificador visual, como o CLIP ou o ViT, diretamente a um LLM central. Essa união permite o alinhamento semântico profundo entre o processamento de imagens, vídeos e textos (LI et

al., 2025). Essa integração multimodal possibilita que o modelo realize tarefas no chamado cenário *zero-shot*, extraindo as capacidades de generalização do CLIP e do LLM para atuar sem qualquer necessidade de treinamento prévio ou específico para a tarefa final (RADFORD et al., 2021). No entanto, o desempenho *zero-shot* em tarefas abstratas, como o reconhecimento de emoções evocadas, ainda carece de validação em larga escala frente à instabilidade contextual e aos vieses inerentes (BHATTACHARYYA; WANG, 2025).

1.6 Ajuste Fino (Fine-Tuning) e Adaptação de Baixo Posto (LoRA)

O *fine-tuning* (ajuste fino) consiste em utilizar um modelo de rede neural previamente treinado em um amplo e genérico conjunto de dados e continuar o seu treinamento em um *dataset* menor e especializado. O objetivo é adaptar o conhecimento global do modelo para uma tarefa específica (RADFORD et al., 2021). No paradigma tradicional (*Full Fine-Tuning*), todos os parâmetros da rede (que nos VLMs modernos ultrapassam a marca dos bilhões) são recalibrados durante o treinamento. Esse processo exige uma quantidade massiva de memória VRAM e poder computacional, inviabilizando a execução em hardwares domésticos ou acadêmicos de menor porte.

Para contornar essa barreira, pesquisadores desenvolveram estratégias de Ajuste Fino Eficiente de Parâmetros (PEFT - *Parameter-Efficient Fine-Tuning*). A técnica mais proeminente e utilizada neste escopo é o *Low-Rank Adaptation* (LoRA), proposta por Hu et al. (HU et al., 2021). O LoRA baseia-se na premissa matemática de que as atualizações de pesos necessárias para adaptar um modelo a uma nova tarefa possuem uma dimensão intrínseca baixa (*low intrinsic rank*). Em vez de atualizar a gigantesca matriz de pesos original (W) do modelo, o LoRA congela completamente essa matriz e injeta na arquitetura duas novas matrizes treináveis e significativamente menores (A e B). A atualização dos pesos (ΔW) é computada pela multiplicação matricial destas duas matrizes reduzidas (BA) (HU et al., 2021).

Geralmente aplicadas aos blocos de atenção (*Attention Blocks*) da arquitetura *Transformer*, essas matrizes de baixo posto reduzem o número de parâmetros treináveis em até 10.000 vezes e diminuem o consumo de memória da GPU em três vezes, mantendo uma qualidade de predição equiparável ao treinamento completo (HU et al., 2021). Uma evolução direta dessa técnica é o QLoRA (*Quantized LoRA*) (DETTMERS et al., 2023), que quantiza os pesos do modelo original congelado para apenas 4-bits, reduzindo drasticamente a pegada de memória do modelo base enquanto treina os adaptadores LoRA em alta precisão. Essa conjunção de técnicas é o que viabiliza a execução local e a especialização de LLMs e VLMs contemporâneos.

1.7 Paradigmas Cognitivos na IA: Sistema 1 e Sistema 2

A evolução das arquiteturas de Inteligência Artificial para o reconhecimento de emoções pode ser compreendida sob a ótica da Teoria dos Processos Duais, consolidada por Daniel Kahneman (KAHNEMAN, 2011). Em sua formulação, a cognição opera através de duas vias distintas: o Sistema 1, que age de forma rápida, automática, intuitiva e primariamente emocional; e o Sistema 2, caracterizado por um pensamento mais lento, analítico, racional e deliberado.

No contexto da Visão Computacional aplicada ao domínio afetivo, as Redes Neurais Convolucionais (CNNs) tradicionais operam em um paradigma estritamente análogo ao Sistema 1. Ao receberem um estímulo visual, essas redes processam as matrizes de

pixels de forma imediata e reflexiva, extraindo características geométricas e mapeando-as diretamente para uma classe de saída (HE et al., 2016). Esse processo é altamente eficiente computacionalmente, mas resulta em uma predição baseada em heurísticas visuais superficiais, sem uma reflexão semântica profunda sobre o contexto da cena (LIMA, 2023).

Em contrapartida, os Modelos de Linguagem Multimodais (VLMs), especialmente quando submetidos a metodologias de raciocínio explícito, introduzem a capacidade de emular o Sistema 2 artificialmente. Através do uso de técnicas como *Chain-of-Thought* (CoT) ou arquiteturas dedicadas de raciocínio lógico (*Reasoning*), o modelo abandona o processamento reativo. Em vez disso, ele passa a decompor a imagem iterativamente, analisando a linguagem corporal, o contexto ambiental e a interação entre os elementos antes de concluir qual emoção está sendo evocada (LI et al., 2025). Embora essa deliberação racional imponha um custo computacional significativamente maior, ela se torna imperativa para resolver ambiguidades afetivas complexas que o reconhecimento automático de padrões não é capaz de discernir.

2 Trabalhos Relacionados

A literatura sobre detecção de emoções evoluiu de abordagens puramente visuais e focadas em fisionomia para modelos multimodais holísticos. No campo das arquiteturas especializadas aplicadas, Lima (LIMA, 2023) desenvolveu uma infraestrutura de software em formato de API baseada em uma rede convolucional customizada via *transfer learning* para executar o mapeamento temporal das emoções de estudantes. A solução integrou a captura assíncrona de expressões faciais a um sistema de juiz online para monitorar o estado afetivo dos alunos durante a resolução de exercícios. Embora o sistema tenha se mostrado funcional e viável, a sua avaliação de desempenho evidenciou gargalos técnicos expressivos na identificação de nuances emocionais complexas. Constatou-se, por exemplo, que a acurácia para a detecção de nojo caiu para 60% em decorrência da sobreposição visual com outras classes afetivas (LIMA, 2023).

Complementarmente, outros estudos focaram na estruturação de redes convolucionais profundas tradicionais voltadas para ambientes educacionais digitais. Por exemplo, Akhand et al. (AKHAND et al., 2021) e Zatarain-Cabada et al. (ZATARAIN-CABADA et al., 2017) avaliaram o emprego de aprendizado transferencial para o reconhecimento de estados básicos em Sistemas de Tutoria Inteligente (STI), demonstrando que o monitoramento provê insumos valiosos para a entrega de *feedbacks* adaptativos aos usuários.

Mais recentemente, o foco de pesquisa deslocou-se para os modelos generativos de larga escala. Bhattacharyya e Wang (BHATTACHARYYA; WANG, 2025) conduziram uma avaliação abrangente sobre o comportamento de VLMs aplicados diretamente ao domínio de emoções evocadas através do *benchmark* EVE[1]. O estudo revelou que, embora os modelos generativos exibam flexibilidade semântica expressiva, eles sofrem com flutuações severas de robustez mediante perturbações na ordenação das etiquetas (*labels*) nos *prompts* (BHATTACHARYYA; WANG, 2025).

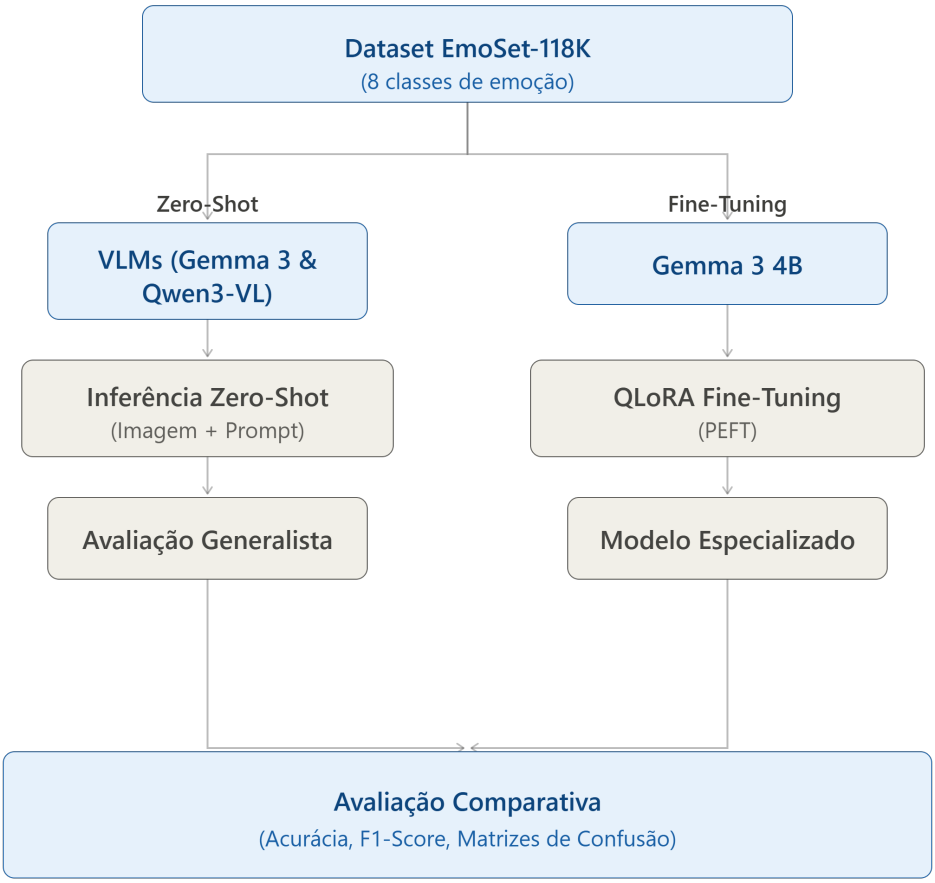
O diferencial deste estudo consiste em expandir as fronteiras de *benchmarks* contemporâneos através de metodologias que contrastam o desempenho natural das IAs (inferência *zero-shot*) com processos complexos de otimização (*fine-tuning*), permitindo discernir os limites de inteligência emocional nativa dos VLMs na predição de cenários de estímulo visual evocativo.

[1] Repositório oficial do *benchmark* EVE disponível em: <https://github.com/rangmiao/eve>

3 Metodologia

Esta pesquisa adota uma abordagem experimental e comparativa. A metodologia proposta visa avaliar a robustez e a precisão de VLMs em tarefas de classificação afetiva perante o cenário evocado.

Figura 1 – Diagrama do fluxo metodológico da pesquisa, evidenciando as abordagens de classificação: *Zero-Shot* Multimodal e Ajuste Fino (LoRA).



Fonte: Elaborado pelo autor.

3.1 Caracterização do *Dataset* EmoSet

Para a condução dos experimentos, foi selecionado o *dataset* EmoSet[2], referenciado na literatura como o primeiro conjunto de dados em larga escala focado em emoções visuais evocadas com anotações ricas de atributos (YANG et al., 2023). O EmoSet difere

[2] Página oficial do *dataset* EmoSet disponível em: <https://vcc.tech/EmoSet>

de bases anteriores em quatro aspectos centrais: escala, riqueza de anotações, diversidade e balanceamento de dados.

Em termos de escala, o conjunto completo (EmoSet-3.3M) compreende 3,3 milhões de imagens coletadas da internet, das quais um subconjunto de 118.102 imagens (EmoSet-118K) foi rigorosamente rotulado por anotadores humanos com auxílio computacional (YANG et al., 2023). Este trabalho utiliza o EmoSet-118K, que se destaca por sua diversidade ao incluir tanto imagens de redes sociais quanto fotografias artísticas, capturando uma vasta gama de estímulos visuais do mundo real.

A taxonomia do *dataset* é fundamentada no modelo psicológico de Mikels (MIKELS et al., 2005), que categoriza as emoções em oito classes discretas. Quatro destas são consideradas emoções positivas (diversão, admiração, contentamento e excitação) e quatro são negativas (raiva, nojo, medo e tristeza). Diferentemente de outros *datasets* históricos de Visão Computacional que apresentam distribuições severamente desiguais, o EmoSet-118K mantém um balanceamento rigoroso: cada categoria contém entre 10.660 e 19.828 imagens. Esse equilíbrio é fundamental para a viabilidade do presente trabalho, mitigando a indução de vieses durante o treinamento (*fine-tuning*) e permitindo avaliações *zero-shot* estatisticamente justas. A disposição visual dessas oito categorias afetivas, bem como exemplos práticos da diversidade fisionômica e contextual contida no *dataset*, pode ser observada na Figura 2.

Figura 2 – Exemplos de estímulos visuais do *dataset* EmoSet mapeados de acordo com as oito classes de emoção de Mikels e a correlação com os seus respectivos atributos.

Fonte: Extraído de Yang et al. (2023).

Além das categorias emocionais primárias, o grande diferencial técnico do EmoSet é a anotação de atributos auxiliares desenhados para reduzir a lacuna afetiva (*affective gap*) e prover interpretabilidade causal. Inspirados por estudos psicológicos, estes atributos são divididos em três níveis de complexidade visual:

- **Baixo Nível:** Brilho (*Brightness*) e Saturação/Coloração (*Colorfulness*).
- **Médio Nível:** Tipo de Cena (*Scene Type*) e Classe de Objeto (*Object Class*).
- **Alto Nível:** Expressão Facial (*Facial Expression*) e Ação Humana (*Human Action*).

É imperativo ressaltar que, para a execução empírica desta pesquisa, **nenhum desses atributos auxiliares foi fornecido como entrada (*input*) aos modelos** (sejam as VLMs em inferência *zero-shot* ou após o *fine-tuning*). A decisão metodológica de omitir esses metadados justifica-se por dois pilares fundamentais:

1. **Capacidade de Inferência Intrínseca:** Pressupõe-se que as arquiteturas de aprendizado profundo devem extrair esses padrões visuais autonomamente da imagem bruta.

2. **Viabilidade em Cenário Real de Produção:** Em um ambiente prático de aplicação, os sistemas de reconhecimento afetivo processam capturas inéditas de forma dinâmica, em contextos nos quais atributos previamente anotados por observadores humanos simplesmente não estariam disponíveis.

Portanto, no contexto deste trabalho, a utilidade da estrutura de atributos do EmoSet (ilustrada na composição matricial da Figura 2) restringe-se puramente ao embasamento teórico e à análise qualitativa *a posteriori*. Essa estrutura viabiliza o estudo sobre *quais* características semânticas e estéticas da imagem exercem maior peso de influência nos acertos e erros perante cada classe emocional evocada.

3.2 Protocolo Experimental

Os experimentos foram modelados em duas configurações principais:

1. **Classificação Multimodal *Zero-Shot*:** Os VLMs recebem as imagens para classificação acompanhadas de um *prompt* estruturado para identificação da emoção evocada em uma das 8 categorias do *dataset*, sem ajustes prévios de pesos ou treinamento específico (RADFORD et al., 2021; BHATTACHARYYA; WANG, 2025).

2. **Ajuste Fino (*Fine-Tuning*) Multimodal:** Especialização da arquitetura interna de um VLM de código aberto por meio de técnicas de *Low-Rank Adaptation* (LoRA), visando adaptar o modelo aos padrões estéticos e afetivos presentes no EmoSet-118K.

4 Desenvolvimento e Experimentos

Esta seção detalha a execução prática da pesquisa, abrangendo desde a seleção dos Modelos de Linguagem Multimodais até a infraestrutura utilizada para inferência e *fine-tuning*. É importante destacar que o trabalho foi estruturado para viabilizar o processamento em larga escala em computadores locais, dadas as restrições de custo e limite de submissões associadas a APIs baseadas em nuvem.

4.1 Seleção de Modelos e Viabilidade Computacional

Para garantir a exequibilidade financeira do projeto perante o volume de 118.102 imagens, selecionaram-se modelos de estado da arte com capacidade multimodal nativa voltados para execução em um computador local. Foram avaliadas e testadas, de maneira equânime e integral, as famílias **Gemma 3** (TEAM et al., 2024) (nas variantes 4B, 12B e 27B) e **Qwen3-VL** (BAI et al., 2023) (nas variantes 4B, 8B e 30B) para a tarefa de classificação *zero-shot*.

Durante a fase de planejamento e testes iniciais, iterou-se também sobre o uso de modelos especializados em raciocínio contínuo (*Reasoning*), como a série Gemma 4 e o Qwen3.6-27B. Contudo, a análise sistêmica de *logs* (campos `reasoning_tokens` e `finish_reason='length'`) revelou que a natureza metodológica do *Chain-of-Thought* exigia que as redes gerassem de 150 a 500 *tokens* exclusivos de reflexão interna sobre os elementos visuais antes de emitirem a predição classificatória objetiva. No modelo

de 27 bilhões de parâmetros, essa carga processual resultou em tempos de inferência de aproximadamente 3 minutos por imagem. Projetando essa restrição computacional para o banco de dados completo do EmoSet, a estimativa do tempo de inferência global superou a marca de 200 dias ininterruptos, tornando os modelos de *Reasoning* computacionalmente inviáveis para o escopo deste projeto.

Em contrapartida, os modelos multimodais de inferência direta de ambas as famílias demonstraram-se plenamente viáveis para o processamento local. Por não embutirem os processos longos de reflexão e se beneficiarem do descarregamento de tensores (*GPU Offload*) diretamente na memória de vídeo, esses modelos apresentaram um tempo de processamento viável para este projeto. As versões de menor escala (4B), por exemplo, atingiram uma média superior a 2 imagens processadas por segundo. Essa decisão de projeto permitiu que a validação experimental fosse estendida com sucesso a todas as seis variantes (Gemma 3 4B/12B/27B e Qwen3-VL 4B/8B/30B) sobre o *dataset* completo, viabilizando o mapeamento comparativo do impacto do tamanho dos parâmetros na acurácia afetiva dentro de um cronograma exequível.

4.2 Processo de Inferência *Zero-Shot* e Tolerância a Falhas

No processo de inferência *zero-shot* com os VLMs, os dados foram particionados seguindo uma estratégia de validação cruzada (*K-Fold*), assegurando o isolamento estatístico correto sobre os vetores de imagens.

A infraestrutura de comunicação entre o *script* de automação e os modelos locais foi estabelecida utilizando a biblioteca `openai`[3] em Python. Embora essa biblioteca seja nativamente projetada para consumir serviços em nuvem, ela permite a customização da rota base da API. Desse modo, as requisições foram redirecionadas para uma porta local (*localhost*) hospedada pelo aplicativo *LM Studio*. Este *software* atua como um servidor que carrega os VLMs na memória da máquina e expõe *endpoints* com formatação e requisições idênticas ao padrão REST da OpenAI, permitindo a integração e a interoperabilidade direta do código.

Para garantir o processamento correto do *payload* multimodal, o pacote de dados foi estruturado contendo o *array* da imagem (codificada no padrão Base64) encapsulado em conjunto com uma instrução textual obrigatória, acionando de forma determinística o mecanismo de atenção (*Attention Mechanism*) da rede sobre os elementos visuais. O *prompt* padronizado submetido aos modelos foi redigido em inglês — visando maximizar o alinhamento semântico com a linguagem primária de pré-treinamento das redes — e estruturado conforme o bloco abaixo para fins de reprodutibilidade:

```
Classify the emotion evoked by this image into one of the
following categories: amusement, anger, awe, contentment,
disgust, excitement, fear, sadness. Reply with only the category
name.
```

Para contornar o risco inerente do processamento ininterrupto sobre vastos volumes de dados, desenvolveu-se também uma rotina de tolerância a falhas. O código foi programado de forma que, ao receber comandos de interrupção no terminal (como o uso do Ctrl+C), o sinal é imediatamente interceptado pelo interpretador; a iteração corrente é assegurada,

[3] Repositório oficial da biblioteca `openai` em Python: <https://github.com/openai/openai-python>

sendo calculado o índice, o *fold* em andamento e as métricas de tempo real, com a subscrição controlada desses vetores em um arquivo JSON. Este mecanismo garantiu que interrupções programadas ou acidentais não causassem corrupção de dados ao longo do processo.

4.3 Treinamento Multimodal e Ajuste Fino (*Fine-Tuning*)

Após a conclusão da etapa de inferência *zero-shot*, iniciou-se o processo de *fine-tuning* utilizando exclusivamente o modelo Gemma 3 (4B-IT) como arquitetura base. Essa etapa teve como objetivo investigar se a especialização direcional de um modelo restrito para atributos emocionais presentes no conjunto de dados poderia superar o desempenho das classificações genéricas produzidas de forma nativa.

O ecossistema de treinamento foi instanciado sob o *Windows Subsystem for Linux* (WSL2), alavancando a API gráfica do *ROCm* (versão 6.0) atrelada à aceleração computacional de uma GPU AMD de alta disponibilidade em VRAM. A dinâmica de treinamento de um VLM exigiu engenharia profunda para resolver os estrangulamentos matemáticos e de memória:

- **Estratégia PEFT/LoRA e Escopo de Hardware:** A escolha metodológica de restringir a etapa de *fine-tuning* exclusivamente à variante base do Gemma 3 (4B-IT) deve-se a limites físicos de infraestrutura. A aplicação do algoritmo sobre matrizes de modelos maiores, como as versões de 12B ou 30B, exigiria dezenas de gigabytes de memória de vídeo, extrapolando os recursos locais de VRAM da GPU AMD utilizada. Para viabilizar a otimização em ambiente *desktop*, parametrizou-se a injeção de adaptadores de baixo *rank* (*LoRA*) (HU et al., 2021). As adaptações LoRA foram aplicadas diretamente sobre projeções específicas dos blocos de atenção do LLM interno (`q_proj`, `v_proj`, `o_proj` e `gate_proj`), restringindo o treinamento a matrizes adicionais de baixa densidade e reduzindo significativamente o consumo de memória e o número de parâmetros treináveis.
- **Tratamento de Colapso Matemático (*Loss Explosion*):** Durante janelas extensas de treinamento (superiores a 11 horas), observou-se a falha crítica das funções de perda, com gradientes gerando valores espúrios e subversivos de `NaN` e `0`. As contra-medidas técnicas exigiram reestruturar a taxa de aprendizado para `2e-5`, impor fases de aquecimento inicial de tensores equivalentes a 5% das etapas de treinamento (*Warmup*), e instanciar travas rígidas via limitação espacial do gradiente (*Gradient Clipping* a `max_grad_norm=1.0`), estabilizando a rede neural.

O ambiente foi resguardado ativamente contra falhas contínuas de estresse térmico em operação. Para mitigar degradações do *hardware* durante as sessões de longas horas, implementou-se o controle vetorial de pico energético (*Power Limit*) do *software* nativo, restringindo a placa de vídeo ao limite operacional seguro de consumo, sem prejuízos práticos substanciais ao fluxo do treinamento. Dessa forma, o resultado final consolida os pesos do LoRA com a base estática para executar a extração e validação do modelo especializado nos dados de teste.

5 Avaliação e Discussão dos Resultados

A avaliação da hipótese central deste trabalho exigiu o mapeamento do potencial gerativo emergente das arquiteturas multimodais contemporâneas e a verificação empírica

de processos de adaptação forçada. Esta seção detalha as métricas obtidas, segmentando a análise entre o estudo comparativo das famílias de VLMs em modo *Zero-Shot* e o impacto estrutural gerado pelo processo de *Fine-Tuning*.

5.1 Desempenho *Zero-Shot* e o Impacto da Escala de Parâmetros

A bateria de testes *zero-shot* submeteu as famílias Gemma 3 e Qwen3-VL a condições de inferência sem nenhum contato prévio ou ajustes nas matrizes de pesos focados no *dataset*. A Tabela 1 sumariza o desempenho global das variantes testadas, introduzindo a métrica de Desvio Padrão (*Std Dev*) do F1-Score no *K-Fold*.

Tabela 1 – Desempenho global e métricas de estabilidade (*K-Fold*) dos VLMs em inferência *Zero-Shot*.

Modelo	**Acurácia Global**	**Macro F1-Score**	**Desvio Padrão (F1)**	**Macro Precisão**
Gemma 3 4B	55,00%	0,4396	0,0313	0,4866
Gemma 3 12B	57,07%	0,5140	0,0232	0,5495
Gemma 3 27B	56,16%	0,5099	0,0256	0,5442
Qwen3-VL 4B	45,99%	0,4019	0,0022	0,5251
Qwen3-VL 8B	49,10%	0,3779	0,0175	0,4721
Qwen3-VL 30B	56,89%	0,4492	0,0188	0,5125

Fonte: Elaborado pelo autor.

O Desvio Padrão é um indicador estatístico relevante nesta análise. Como o modelo atua no cenário *zero-shot*, seus parâmetros e conhecimentos estão integralmente congelados durante os testes. Embora a decodificação generativa possua leve estocasticidade em função da *seed* aleatória inerente a cada requisição, a rede não sofre qualquer atualização de pesos ou aprendizado entre uma dobra e outra. Portanto, a variância dos resultados observada entre as partições (*folds*) não indica uma instabilidade interna da arquitetura, mas reflete, primordialmente, a heterogeneidade orgânica dos dados contidos em cada *fold*, sinalizando que determinados subconjuntos apresentam estímulos visuais de interpretação mais fácil ou mais ambígua.

Os resultados indicam uma notável capacidade de compreensão semântica por parte dos modelos generativos, dadas as condições puramente *zero-shot*. Para mapear o espectro analítico, a Tabela 2 exibe o detalhamento das predições de acurácia por classe de emoção evocada para todos os VLMs avaliados.

Tabela 2 – Acurácia individual por classe para todas as variantes dos modelos Gemma 3 e Qwen3-VL (*Zero-Shot*).

Classe Emocional	**Gemma 3 4B**	**Gemma 3 12B**	**Gemma 3 27B**	**Qwen3-VL 4B**	**Qwen3-VL 8B**	**Qwen3-VL 30B**
Amusement	16,60%	23,87%	32,56%	38,29%	19,95%	28,53%
Anger	40,09%	39,25%	52,79%	28,53%	30,09%	19,00%
Awe	57,35%	69,05%	47,80%	33,86%	72,56%	53,59%
Contentment	**84,66%**	72,48%	71,43%	74,07%	75,82%	76,10%
Disgust	54,84%	62,95%	51,32%	19,02%	23,13%	73,06%
Excitement	71,03%	67,06%	65,61%	57,02%	63,32%	78,26%
Fear	41,82%	49,64%	49,71%	34,84%	29,72%	51,82%
Sadness	74,45%	76,23%	**81,55%**	68,01%	67,76%	69,79%

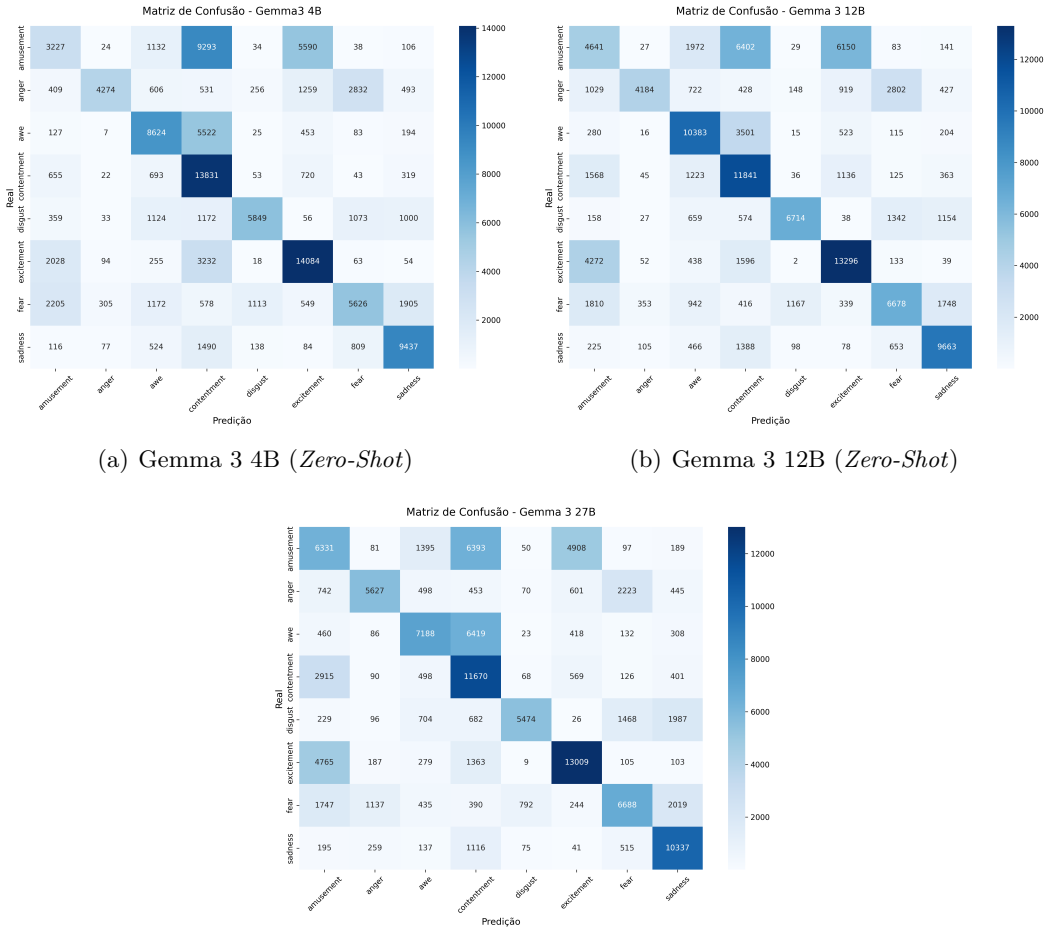
Fonte: Elaborado pelo autor.

A Figura 3 ilustra o comportamento das diferentes variantes da família Gemma 3 no cenário *zero-shot*. Observa-se que a variante de 12B apresentou o melhor desempenho global da família (57,07%), superando inclusive a versão de 27B (56,16%). Esse resultado

sugere que o aumento do número de parâmetros não implicou, necessariamente, em maior capacidade de discriminação da emoção evocada de forma generalista. Uma possível explicação é que modelos maiores tendem a priorizar abstrações semânticas textuais mais amplas, enquanto o reconhecimento de emoções focadas depende fortemente de pistas visuais e interpretações psicológicas intrincadas.

Além disso, as três variantes exibiram padrões de erro semelhantes, particularmente na classe *Amusement*, que apresentou as menores taxas de acerto da família, variando entre 16,60% e 32,56%. As matrizes de confusão indicam que essa classe foi frequentemente confundida com *Contentment* e *Excitement*, possivelmente em razão da proximidade semântica entre emoções de valência positiva e das sobreposições contextuais presentes no conjunto de dados.

Figura 3 – Matrizes de Confusão da família Gemma 3 em modo *Zero-Shot*. É notável o colapso semântico contínuo na discriminação fina da classe *Amusement*.



(a) Gemma 3 4B (*Zero-Shot*)

(b) Gemma 3 12B (*Zero-Shot*)

(c) Gemma 3 27B (*Zero-Shot*)

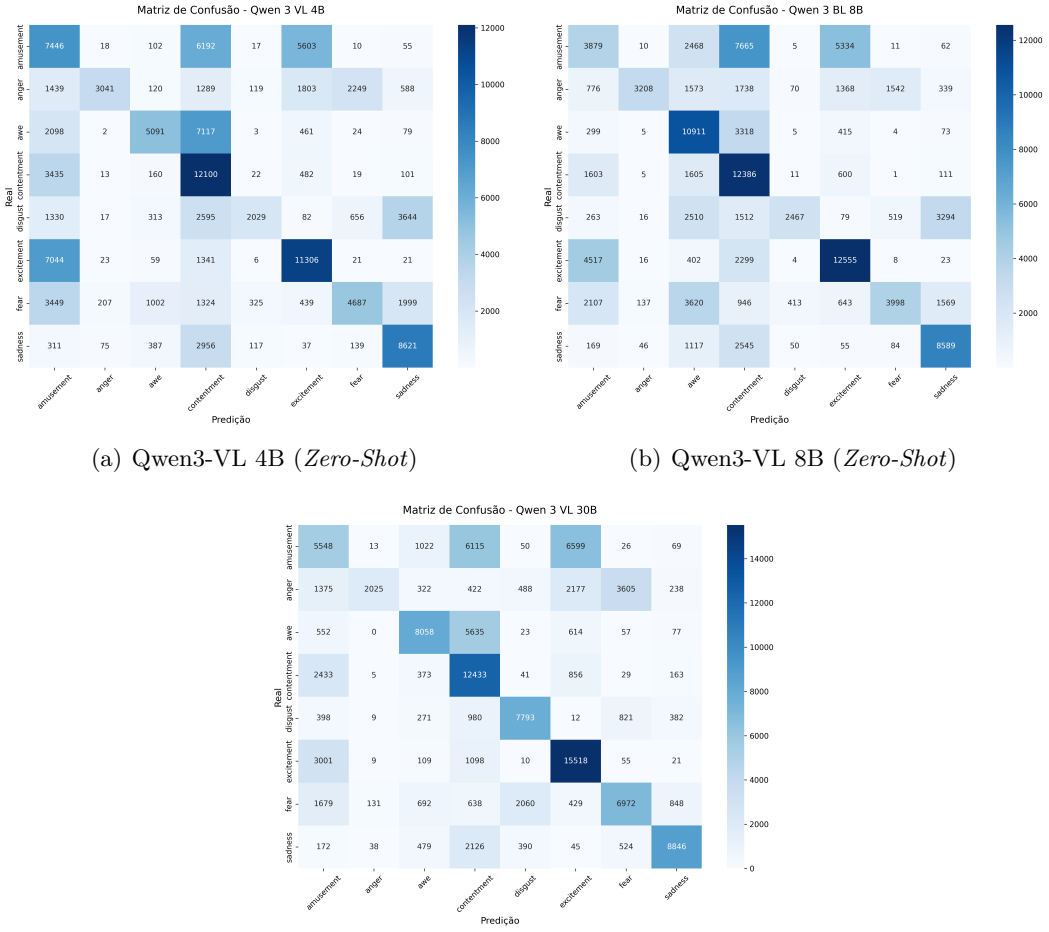
Fonte: Elaborado pelo autor.

Por sua vez, a família Qwen3-VL (Figura 4) exibiu uma correlação mais direta entre escala e desempenho (de 45,99% no modelo 4B para 56,89% no 30B). A versão de 4B demonstrou forte instabilidade na classe *Disgust* (apenas 19,02% de acertos, vide Tabela 2). À medida que o modelo escala para 30B, a rede adquire acuidade estrutural para isolar o

Nojo (atingindo 73,06%), revelando que parâmetros adicionais expandem a granularidade analítica nativa do modelo para percepções contextuais mais nocivas.

Ao considerar os atributos multivariados do EmoSet (detalhados na Seção 3), é possível interpretar qualitativamente a origem dos colapsos em classes específicas. A baixa eficiência em *Amusement* e *Contentment* observada em ambas as famílias, por exemplo, pode estar associada à sobreposição de atributos de Nível Baixo (alta Saturação e Colorfulness) e Nível Médio (*Scene Types* envolvendo atividades recreativas). Como os VLMs em regime *zero-shot* analisam o tom geral da cena, diante de ambientes vibrantes e ensolarados, as redes assumem uma valência global positiva, sendo incapazes de inferir unicamente através do raciocínio reflexivo rápido se a atmosfera da cena dita um relaxamento passivo (*Contentment*) ou um entretenimento ativo (*Amusement*).

Figura 4 – Matrizes de Confusão da família Qwen3-VL em modo *Zero-Shot*. Observa-se a melhora gradativa da IA generativa nas classes de *Disgust* e *Excitement* com o aumento de parâmetros.



(a) Qwen3-VL 4B (*Zero-Shot*)

(b) Qwen3-VL 8B (*Zero-Shot*)

(c) Qwen3-VL 30B (*Zero-Shot*)

Fonte: Elaborado pelo autor.

5.2 Análise Qualitativa do Ajuste Fino: Colapso de Classe e Atalhos

A tentativa de especializar a arquitetura por meio de *fine-tuning* quantizado (QLoRA) (DETTMERS et al., 2023) restrito ao Gemma 3 4B revelou um paradoxo analítico marcante. O exame isolado das métricas globais poderia sugerir um progresso na especialização gerativa. No entanto, a Tabela 3 contrapõe os resultados antes e depois do ajuste do VLM, evidenciando um cenário estrito de Colapso de Classe, onde a rede generativa colapsou seu espaço latente textual-visual durante a adaptação.

Tabela 3 – Comparativo detalhado de métricas: Gemma 3 4B (*Zero-Shot*) versus Gemma 3 4B (*Fine-Tuning*).

Métrica / Classe	**Gemma 3 4B (*Zero-Shot*)**	**Gemma 3 4B (*Fine-Tuning*)**
Acurácia Global	55,00%	56,27%
Macro F1-Score	0,4396	0,5287
Desvio Padrão F1	0,0313	0,0014
Amusement	16,60%	**0,83%**
Anger	40,09%	49,19%
Awe	57,35%	**85,14%**
Contentment	84,66%	69,52%
Disgust	54,84%	**88,91%**
Excitement	71,03%	81,93%
Fear	41,82%	44,53%
Sadness	74,45%	**40,80%**

Fonte: Elaborado pelo autor.

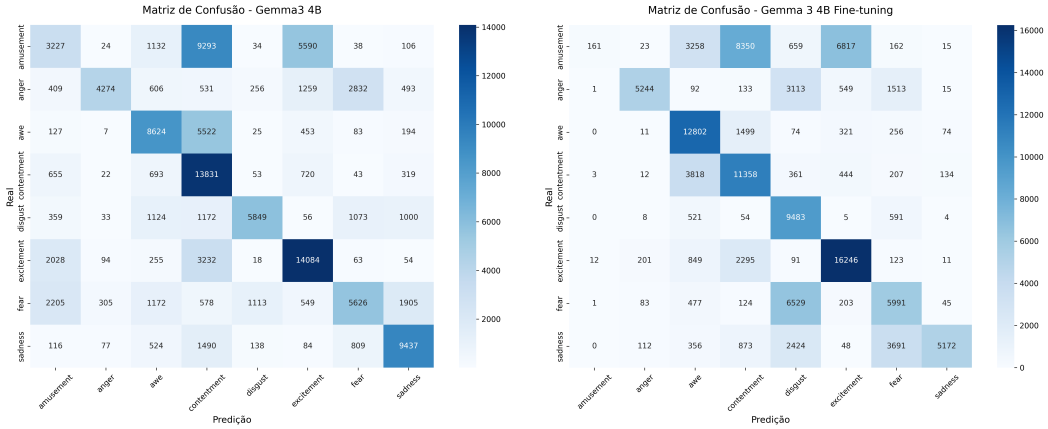
Esse resultado aponta a ocorrência severa do fenômeno de aprendizado de atalhos (*shortcut learning*), no qual a atualização dos pesos do VLM passou a priorizar heurísticas visuais contrastantes presentes no *dataset*, em detrimento das representações semânticas flexíveis adquiridas durante o pré-treinamento global. Consequentemente, a minimização da função de perda gerou um viés direcional que inflou artificialmente as predições para as classes Admiração (*Awe* - 85,14%) e Nojo (*Disgust* - 88,91%).

O custo prático dessa convergência foi a destruição total da granularidade preditiva fina. A precisão da predição da classe *Amusement* diminuiu consideravelmente, caindo de 16,60% para insignificantes 0,83% de acertos. A observação do quadrante inferior direito da Figura 5b indica uma distorção paralela nas emoções negativas: o modelo modificado perdeu a capacidade original de diferenciar a Tristeza (*Sadness* - cuja acurácia caiu abruptamente de 74,45% para 40,80%), classificando falhas semânticas de medo e tristeza indiscriminadamente como Nojo (*Disgust*). Sob um regime de treinamento singular, os adaptadores LoRA priorizaram estatísticas enviesadas do *dataset* e comprometeram a fidelidade subjetiva da avaliação gerativa.

Conclusões e Trabalhos Futuros

O presente trabalho avaliou o potencial gerativo emergente dos Modelos de Linguagem Multimodais (VLMs) atuando no reconhecimento de emoções evocadas, contrastando abordagens de inferência direta e de otimização. Os resultados dos experimentos com arquiteturas *zero-shot* sugerem que VLMs podem apresentar mecanismos interessantes de identificação semântica abstrata do contexto sem a necessidade de treinamento supervisi-

Figura 5 – Estudo de impacto do *fine-tuning* quantizado no modelo Gemma 3 4B, demonstrando graficamente a supressão de classes minoritárias após o sobreajuste adaptativo.



(a) Gemma 3 4B (*Zero-Shot* - Conhecimento Generalista)

(b) Gemma 3 4B (*Fine-Tuning* - Heurísticas e Colapso)

Fonte: Elaborado pelo autor.

onado estrito. O experimento focado em *fine-tuning* de um VLM base demonstrou que, sob os hiperparâmetros de eficiência e as restrições de *hardware* impostas neste estudo, o processo de adaptação contínua resultou em convergência catastrófica e colapso de classe associado ao *shortcut learning*. Esse cenário indica que a falha não é necessariamente uma fraqueza orgânica irrevogável da arquitetura, mas comprova que a especialização direcional de VLMs sobre dados puramente afetivos exige otimizações hiperparamétricas muito mais rigorosas para evitar o sequestro do *prompting* flexível.

Além dos resultados quantitativos obtidos, este trabalho contribuiu para a discussão acadêmica sobre os limites e potencialidades de arquiteturas generativas em tarefas focadas de reconhecimento emocional evocado. A pesquisa evidenciou que, embora VLMs possuam viabilidade na percepção semântica da composição de imagem em regime *zero-shot*, ainda apresentam fragilidades substanciais na adaptação restrita a domínios emocionalmente ambíguos sem comprometer o seu espaço latente primordial.

Para dar continuidade a esta linha de pesquisa e mitigar as falhas metodológicas de generalização e estabilidade observadas, recomendam-se as seguintes explorações para trabalhos futuros:

- **Implementação de *Chain-of-Thought* (CoT):** Em experimentos futuros, o *prompt* poderá instruir o modelo a descrever inicialmente características de ambiente, contextuais e estéticas da cena para, apenas posteriormente de forma induzida, inferir a classe da emoção evocada correspondente. Essa estratégia pode estimular o encadeamento das etapas intermediárias, potencialmente ampliando a capacidade analítica limpa dos transformadores.
- **Adoção de Modelos Focados em Raciocínio (System 2):** Avaliar VLMs projetados explicitamente para *reasoning* avançado que consigam decompor computacio-

nalmente a reflexão de imagens visuais complexas, mitigando ambiguidades subjetivas e generalizações precipitadas do Sistema 1 de forma arquitetural nativa.

- **Avaliação *Few-Shot*:** Testar IAs generativas com *prompts* carregados em contexto alimentados com 3 a 5 amostras resolvidas (*few-shot*), calibrando provisoriamente na janela de contexto a associação de diretrizes de anotação específicas do *dataset* evocado sem forçar adaptações definitivas de parâmetros de peso (*LoRA*).
- **Otimização Matemática da Função de Perda:** No processo de *fine-tuning*, substituir a entropia cruzada tradicional por técnicas consolidadas e direcionais contra desbalanceamento semântico como a *Focal Loss* ou mineração de falsos negativos rigorosos (*Hard Negative Mining*), protegendo a absorção dos metadados finos da imagem para classes sutis contra o colapso e o engolimento provocado pelas super-categorias visuais.
- **Classificação Bidimensional Evocada:** Modificar o arranjo de validação para medir e mapear diretamente não as classes estritas pré-rotuladas, mas os sub-vetores psicológicos do modelo circunplexo subjacente à cena em análise. Essa classificação multi-etapa toleraria de forma superior a subjetividade intrínseca da resposta e percepção humana, assumindo formalmente que estímulos visuais complexos não evocam valências emocionais unidimensionais e exclusivas.

Referências

AKHAND, M. et al. Facial emotion recognition using transfer learning in the deep cnn. *Electronics*, v. 10, n. 9, p. 1036, 2021. Citado 3 vezes nas páginas 2, 3 e 6.

BAI, J. et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. Disponível em: <https://arxiv.org/abs/2308.12966>. Acesso em: 10 jul. 2026. Citado na página 9.

BHATTACHARYYA, S.; WANG, J. Z. Evaluating vision-language models for emotion recognition. *arXiv preprint arXiv:2502.05660v1*, 2025. Disponível em: <https://arxiv.org/abs/2502.05660>. Acesso em: 10 jul. 2026. Citado 6 vezes nas páginas 2, 3, 4, 5, 6 e 9.

DETTMERS, T. et al. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, v. 36, 2023. Citado 2 vezes nas páginas 5 e 15.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2021. Citado na página 4.

EKMAN, P. Facial expression and emotion. *American psychologist*, American Psychological Association, v. 48, n. 4, p. 384, 1993. Citado na página 3.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778. Citado 2 vezes nas páginas 3 e 6.

HU, E. J. et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. Disponível em: <https://arxiv.org/abs/2106.09685>. Acesso em: 10 jul. 2026. Citado 2 vezes nas páginas 5 e 11.

KAHNEMAN, D. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux, 2011. Citado na página 5.

LI, Z. et al. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*, 2025. Disponível em: <https://arxiv.org/abs/2501.02189>. Acesso em: 10 jul. 2026. Citado 3 vezes nas páginas 2, 5 e 6.

LIMA, P. V. B. O. *Detecção Automática de Emoções no Aprendizado de Algoritmos*. Monografia (Bacharelado em Ciência da Computação) — Universidade Federal do Maranhão, São Luís, 2023. Citado 3 vezes nas páginas 2, 3 e 6.

MIKELS, J. A. et al. Emotional category data on images from the international affective picture system. *Behavior research methods*, Springer, v. 37, n. 4, p. 626–630, 2005. Citado na página 8.

RADFORD, A. et al. Learning transferable visual models from natural language supervision. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2021. p. 8748–8763. Citado 4 vezes nas páginas 2, 4, 5 e 9.

TEAM, G. et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. Disponível em: <https://arxiv.org/abs/2403.08295>. Acesso em: 10 jul. 2026. Citado na página 9.

VASWANI, A. et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2017. v. 30. Citado na página 4.

YANG, J. et al. Emoset: A large-scale visual emotion dataset with rich attributes. *arXiv preprint arXiv:2307.07961v2*, 2023. Disponível em: <https://arxiv.org/abs/2307.07961>. Acesso em: 10 jul. 2026. Citado 3 vezes nas páginas 4, 7 e 8.

ZATARAIN-CABADA, R. et al. Building a face expression recognizer and a face expression database for an intelligent tutoring system. In: IEEE. *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. [S.l.], 2017. p. 391–393. Citado 2 vezes nas páginas 2 e 6.

MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
DEPARTAMENTO DE COMPUTAÇÃO - NG

FOLHA DE APROVAÇÃO DE TCC Nº 3 / 2026 - DECOM (11.56.03)

Nº do Protocolo: 23062.031441/2026-71

Belo Horizonte-MG, 18 de junho de 2026.

GUSTAVO CAMPOS PÁDUA

AVALIAÇÃO DA CAPACIDADE DE DETECÇÃO DE EMOÇÕES EM MODELOS DE LINGUAGEM MULTIMODAIS (VLMS)

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Computação do Centro Federal de Educação Tecnológica de Minas Gerais, do Campus Nova Gameleira, como parte do requisito para obtenção do grau de Engenheiro(a) de Computação

Aprovado em 18 de junho de 2026.

Banca Examinadora:

Prof. (D.Sc.) Ismael Santana Silva - Orientador
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. (D.Sc.) Danilo Boechat Seufitelli
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. (D.Sc.) Fábio Rocha da Silva
Centro Federal de Educação Tecnológica de Minas Gerais

Belo Horizonte
2026

(Assinado digitalmente em 18/06/2026 15:48)
DANILO BOECHAT SEUFITELLI
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 1371545

(Assinado digitalmente em 19/06/2026 15:09)
FABIO ROCHA DA SILVA
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 1703246

(Assinado digitalmente em 25/06/2026 15:34)
ISMAEL SANTANA SILVA
PROFESSOR ENS BASICO TECN TECNOLOGICO
DECOM (11.56.03)
Matrícula: 1028758

Visualize o documento original em **https://sig.cefetmg.br/public/documentos/index.jsp** informando seu número: **3**, ano: **2026**, tipo: **FOLHA DE APROVAÇÃO DE TCC**, data de emissão: **18/06/2026** e o código de verificação: **c7efca22d7**