UNIVERSIDADE FEDERAL DE GOIÁS / INSTITUTO DE INFORMÁTICA

Multimodal Large Language Models para Detecção de Emoções

Avaliação de Estratégias de Prompting

Fernanda Bufon Farber

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

UNIVERSIDADE FEDERAL DE GOIÁS (UFG)
INSTITUTO DE INFORMÁTICA (INF)

FERNANDA BUFON FARBER

Multimodal Large Language Models para Detecção de Emoções

Avaliação de Estratégias de Prompting

Goiânia
2025

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): FERNANDA BUFON FARBER

Título do trabalho: Multimodal Large Language Models para Detecção de Emoções

Avaliação de Estratégias de Prompting

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Fernanda Bufon Farber**, **Usuário Externo**, em 04/02/2026, às 19:31, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fernando Marques Federson**, **Professor do Magistério Superior**, em 13/03/2026, às 11:30, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

FERNANDA BUFON FARBER

Multimodal Large Language Models para Detecção de Emoções

Avaliação de Estratégias de Prompting

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Orientador: Prof. Dr. Fernando Marques Federson

Goiânia

2025

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UFG.

FARBER, FERNANDA BUFON
Multimodal Large Language Models para Detecção de Emoções [manuscrito]: Avaliação de Estratégias de Prompting / FERNANDA BUFON FARBER. - 2025.
64 f.: 2025

Orientador: Prof. Dr. Fernando Marques Federson
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Informática (INF), Inteligência Artificial, Goiânia, 2025.

1. Inteligência Artificial. 2. Multimodal Large Language Models. 3. Reconhecimento de Emoções.

I. Federson, Fernando Marques , orient. II. Título.

CDU 004

FERNANDA BUFON FARBER

Multimodal Large Language Models para Detecção de Emoções

Avaliação de Estratégias de Prompting

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Data da Aprovação: 09 de dezembro de 2025.

Prof. Dr. Fernando Marques Federson
Orientador (INF-UFG)

Prof. Dr. Aldo André Díaz Salazar
Coordenador de TCC do BIA (INF-UFG)

Prof. Dr. Anderson da Silva Soares
Coordenador do BIA (INF-UFG)

Prof. Dr. Alexandre Gomes de Siqueira
(University of Florida)

FERNANDA BUFON FARBER

Multimodal Large Language Models para Detecção de Emoções

Avaliação de Estratégias de Prompting

RESUMO

Este Relatório de Conclusão de Curso tem como objetivo reunir os resultados da minha jornada para me tornar um especialista em **Multimodal Large Language Models**. Uma ilustração e sua narrativa descrevem os períodos de trabalho. Os Apêndices contêm os Termos de Aceite de Entrega e os resultados obtidos durante cada período de trabalho.

Palavras-chave: Inteligência artificial; Multimodal large language models; Reconhecimento de emoções.

ABSTRACT

This Course Completion Report aims to bring together the results of my journey to become an expert in **Multimodal Large Language Models**. An illustration and its narrative describe the work periods. The Appendices contain the Delivery Acceptance Terms and the results obtained during each work period.

Keywords: Artificial intelligence; Multimodal large language models; Emotion recognition.

Goiânia

2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

Minha Jornada

Fernanda Bufon Farber

Especialista em: Multimodal Large Language Models

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

MINHA JORNADA

Nome: **Fernanda Bufon Farber**

Especialidade: **Multimodal Large Language Models**

Objetivo deste documento

Durante o processo da disciplina Residência em IA[1], foram gerados diversos resultados na construção da minha especialização. A cada semana, um conjunto de resultados foi formalizado por um Termo de Aceite de Entrega e avaliado por uma banca, considerando o planejado e o realizado para o período. Este documento tem como objetivo descrever esses resultados obtidos, fazendo referência aos Termos de Aceite de Entrega e seus documentos associados.

Minha Jornada

Minha Jornada se iniciou na **Semanas 1** com a escolha da área na qual eu gostaria de me aprofundar: Inteligência Artificial Multimodal (IA Multimodal). Optei por seguir esse caminho porque se trata de uma área que tem avançado rapidamente e vem sendo muito utilizada em pesquisas e aplicações de IA. Além disso, o curso de Inteligência Artificial da Universidade Federal de Goiás não oferece uma disciplina específica sobre o tema e também não existe um grupo de estudo ou interesse relacionado. Sempre tive vontade de explorar esse assunto com mais profundidade, mas nunca havia encontrado uma oportunidade adequada. Por ser meu primeiro contato com a área, adotei uma estratégia simples: anotei em um caderno tudo o que eu já sabia sobre *Multimodal Models* (Modelos Multimodais), de forma clara e direta, e em seguida listei todas as dúvidas que eu tinha sobre o tema. Isso me ajudou a direcionar melhor o aprendizado. Depois dessa etapa inicial, comecei a buscar conteúdos que me ajudassem a entender melhor o que são Modelos Multimodais, como surgiram, como são utilizados e em quais contextos se aplicam. Para isso, reuni artigos, vídeos e aulas que considerei relevantes e organizei também uma lista de

[1] Dez Semanas, entre setembro de 2025 e dezembro de 2025.

prioridades para otimizar o tempo de estudo. Mais detalhes sobre as referências consultadas e as atividades realizadas nesse período estão descritos no **Apêndice 1**.

Já na **Semana 2**, comecei de fato o estudo aprofundado em IA Multimodal, avançando na leitura dos artigos e acompanhando as aulas selecionadas. Esse contato inicial foi essencial para o meu direcionamento, pois me permitiu concluir que meu interesse não estava apenas nos Modelos Multimodais de forma geral, mas especificamente nos MLLMs (*Multimodal Large Language Models*), um subconjunto que integra diferentes modalidades a partir de um LLM treinado em linguagem. Para mais detalhes, consulte o **Apêndice 2.** Perceber essa distinção foi um ponto importante na minha jornada, pois me fez compreender que a multimodalidade é um campo amplo, com os MLLMs representando uma vertente mais específica e na qual eu desejo continuar me aprofundando.

A partir do melhor entendimento do tema e da definição da vertente específica que eu desejava explorar, nas **Semanas 3 e 4** comecei a me aprofundar mais na área de *Multimodal Large Language Models*. Nessa fase, busquei compreender com mais precisão o funcionamento desses modelos, selecionando artigos atuais para analisar as arquiteturas multimodais recentes e mais utilizadas. Foi também nesse período que comecei a enxergar um possível caminho de investigação: percebi que a interação com o usuário pode se tornar mais empática quando o LLM é capaz de interpretar emoções, sobretudo a partir de informações visuais. Com isso, minha ideia de pesquisa começou a “ganhar forma”. Mais detalhes sobre o que foi feito podem ser conferidos no **Apêndice 3.** Além disso, elaborei uma lista com os modelos multimodais mais usados atualmente (abertos ou não) como Qwen3-VL, LLaVA, entre outros.

A **Semana 4** marcou um aprofundamento ainda maior no estudo das arquiteturas. Utilizando a ferramenta NotebookLM, consolidei as informações presentes nos artigos encontrados na **Semana** anterior e gerei diversos conteúdos didáticos a partir deles, como: vídeos, podcasts e mapas mentais. Também avancei da pesquisa puramente teórica para a organização prática dos testes que eu planejava executar. Para isso, construí uma planilha estruturada contendo os diferentes modelos multimodais disponíveis e suas formas de

acesso (API, código, interface), a qual está registrada no **Apêndice 4**. Durante essa etapa, analisei também um artigo que propõe uma estratégia multimodal para entendimento de vídeos com humanos e detecção de emoções, o que reforçou a percepção de que há um interesse crescente da comunidade científica em trabalhar nesse tema.

As **Semanas 5 e 6** foram utilizadas para a consolidação técnica e o início dos experimentos. Na **Semana 5**, especificamente, o foco foi entender o desempenho dos MLLMs no contexto de detecção de emoção. Para isso, executei testes práticos com três modelos multimodais que eu havia listado anteriormente (todos da família GPT), o que me permitiu compreender melhor como utilizá-los e avaliar seu comportamento inicial. Paralelamente, realizei uma pesquisa exploratória para encontrar artigos que testaram esses mesmos modelos, buscando identificar quais MLLMs seriam mais adequados para o meu contexto de uso e avaliando prós e contras de diferentes alternativas. Mais detalhes sobre a pesquisa podem ser encontrados no **Apêndice 5.**

Com os modelos selecionados, pude sistematizar melhor o processo de testes na **Semana 6**. Defini dois datasets, estabeleci a quantidade de dados por categoria emocional e determinei critérios de avaliação para manter consistência entre os experimentos. O documento que descreve esse processo está disponível no **Apêndice 6.** Para garantir versionamento e organização adequada do projeto, também criei um repositório no GitHub e uma organização no Hugging Face para armazenar os dados e códigos utilizados. Além disso, avancei na familiarização com os modelos selecionados, aprendendo quais ferramentas e frameworks usar e como acessá-los de forma eficiente.

Nas **Semanas 7 e 8**, precisei tomar uma decisão importante. Apesar de já ter sistematizado meus testes, percebi que era necessário ampliar significativamente a base de dados, pois eu havia conseguido apenas cerca de 100 amostras, o que seria insuficiente para chegar a conclusões consistentes. Além disso, notei que eu ainda precisava me aprofundar mais no tema de detecção de emoção com modelos multimodais antes de tentar realizar experimentos sem seguir uma *baseline* sólida. Era fundamental responder questões como: quais modelos os artigos utilizam para essa tarefa, quais *datasets* são mais usados,

quais métodos são aplicados e qual é, de fato, o desempenho dos MLLMs nesse contexto. Assim, na **Semana 7**, comecei a reunir e filtrar artigos relacionados à aplicação de MLLMs em detecção de sentimentos, realizando uma pesquisa mais profunda em diferentes congressos, lendo *abstracts* e selecionando os trabalhos mais relevantes para um estudo detalhado (mais informações estão no **Apêndice 7**). O objetivo principal dessa etapa era encontrar um trabalho viável para realizar uma replicação experimental, ou seja, um trabalho cujos métodos, cenários e resultados fossem realisticamente replicáveis.

Com isso em mente, na **Semana 8**, selecionei o trabalho de Fang et al.[2], que inicialmente pareceu uma boa opção por não exigir treinamento ou fine-tuning. No entanto, após a leitura, senti falta de confiança nos métodos utilizados, além de apresentar apenas uma pequena melhora no desempenho dos MLLMs na tarefa de detecção de emoção (detalhes no **Apêndice 8**). Essa constatação reforçou ainda mais a importância de ampliar a busca e realizar uma filtragem cuidadosa para encontrar um artigo que realmente valesse a pena replicar e que apresentasse resultados consistentes e relevantes para minha pesquisa.

A partir dessa necessidade, na **Semana 9,** realizei uma busca extensiva para encontrar o artigo. Para mais detalhes, veja **Apêndice 9**. Após analisar todos os artigos, selecionei o trabalho de Wang et al.[3], que se mostrou inicialmente viável para replicação. Paralelamente à leitura desse artigo, desenvolvi o código necessário para realizar chamadas a um modelo MLLM enviando simultaneamente uma mensagem e uma imagem, permitindo que o modelo retornasse um texto com o resultado da predição. A partir do artigo e da minha versão inicial de detecção de sentimento implementada, organizei como seriam conduzidos os experimentos, definindo os passos e adaptações necessárias. Além disso, comecei a seguir e replicar o *pipeline* proposto pelos autores do artigo, levando em consideração as minhas limitações computacionais e o ambiente disponível para execução.

[2] FANG, Yiyang; LIANG, Jian; HUANG, Wenke; LI, He; SU, Kehua; YE, Mang. *Catch Your Emotion: Sharpening Emotion Perception in Multimodal Large Language Models*. In: Proceedings of the 42nd International Conference on Machine Learning – ICML 2025. PMLR 267, Vancouver, 2025.

[3] WANG, Zhifeng; ZHANG, Qixuan; ZHANG, Peter; NIU, Wenjia; ZHANG, Kaihao; SANKARANARAYANA, Ramesh; CALDWELL, Sabrina; GEDEON, Tom. *Visual and Textual Prompts in VLLMs for Enhancing Emotion Recognition*. arXiv preprint, arXiv:2504.17224v3, 2025.

Na **Semana 10,** consegui realizar a replicação do trabalho em uma escala menor e gerei os primeiros resultados. Os valores obtidos ficaram abaixo do que foi apresentado no artigo original, especialmente quando analisados de forma comparativa entre o método-base e o método proposto, mas essa diferença era esperada devido a diversos fatores. Entre eles, destaco o fato de não ter utilizado exatamente o mesmo modelo empregado pelos autores, por limitações de acesso ao framework original, além das diferenças nos dados utilizados (tanto em tipo quanto em quantidade) entre o experimento e a minha replicação (mais detalhes no **Apêndice 10**). Ainda assim, mesmo com resultados diferentes, a comparação entre os métodos, ainda que em um cenário mais controlado e com menos dados, foi extremamente interessante e trouxe alguns insights valiosos. Primeiro, os modelos mantiveram desempenho acima do acaso, indicando que capturam algum padrão, mesmo em um cenário muito limitado. Além disso, os resultados mostraram que abordagens simples, como apenas enviar a imagem e um prompt genérico, tiveram desempenho superior aos métodos mais complexos, incluindo o método do estudo replicado. Isso sugere que, para os modelos utilizados, prompts diretos funcionam melhor do que cadeias de processamento mais elaboradas. Outro ponto importante é que a *pipeline* utilizada por Wang et al. não obteve um bom resultado neste contexto, reforçando que a eficácia desse método depende fortemente do modelo utilizado e das características do dataset.

Com tudo o que vivi nessa Jornada de 10 **Semanas**, concluí que, apesar de o método analisado fazer sentido, ele nem sempre será a melhor escolha, e que a detecção de emoção em modelos multimodais ainda é uma área pouco explorada e com os modelos demonstrando inúmeras dificuldades na tarefa. Esse processo despertou ainda mais meu interesse em investigar o uso de MLLMs como detectores de emoções, especialmente porque essa capacidade pode abrir caminhos para interações mais personalizadas e “reais” com assistentes inteligentes. As limitações que observei me instigaram a entender melhor por que essa tarefa é tão desafiadora, quais classes são mais difíceis para os modelos e como esses padrões variam entre diferentes arquiteturas e famílias de modelos. Essa experiência certamente abre portas para várias pesquisas e investigações futuras dentro da área.

Essa Jornada da Residência, registrada neste Trabalho de Conclusão de Curso, não seria possível sem as pessoas especiais que caminharam comigo. Por isso, gostaria de agradecer profundamente ao professor Fernando Marques Federson, por toda a compreensão, apoio e cuidado ao longo dessa trajetória – não apenas durante a Residência em IA, mas em todo o Bacharelado. Agradeço também aos professores Leonardo Antônio Alves e Cedric Luiz de Carvalho pelas dicas, pela participação nas bancas e pelo apoio constante durante o processo. Estendo meus agradecimentos a todos os professores que, de alguma forma, foram fundamentais para a minha formação acadêmica e contribuíram para que eu chegasse até aqui, em especial aos professores Sávio Teles, Anderson Soares, Telma Woerle, Arlindo Galvão e Erika Morais. Sou imensamente grata aos meus colegas da turma de 2022 do Bacharelado em Inteligência Artificial; o apoio mútuo, a colaboração e o sentimento de estarmos juntos nessa caminhada tornaram a jornada mais leve, significativa e prazerosa. Por fim, deixo um agradecimento especial ao meu namorado e à minha família, que sempre acreditaram em mim, torceram por mim e foram a base das minhas conquistas acadêmicas, profissionais e pessoais.

APÊNDICE 1

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 3 de set. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

Fernanda Bufon Farber

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Primeiro contato com o tema Inteligência Artificial Multimodal:

- Registro inicial do que já sabia sobre Multimodal Models em uma página do caderno, listando tudo que sei sobre o tema em termos simples;
- Levantamento de dúvidas sobre o tema para guiar o aprendizado do fundamento;
- Pesquisa de conteúdos úteis:
    - Pesquisa de papers no Semantic Scholar e no Google Scholar com as queries "Survey Multimodal" e "Multimodal Models":
        - O objetivo é entender a história e os fundamentos principais sobre o tema, além de entender também o cenário atual do mesmo;
        - Nesta etapa, separei 12 papers com base no ano de publicação, na relevância, no número de citações e através da leitura do abstract.
    - Pesquisa de aulas no YouTube com a query "multimodal models class", para que a pesquisa me apresentasse AULAS sobre o tema, e não apenas vídeos explicativos:
        - O objetivo é ter disponível um outro método de aprendizado além da leitura;
        - Nesta etapa, analisei e separei 2 playlists e 6 vídeos, com base no ano de lançamento, número de views e duração da aula.
- Criação de uma lista de prioridades de conteúdos, para otimizar o tempo e não ultrapassar o período de uma Semana nos fundamentos:
    - A prioridade foi definida de acordo com o assunto abordado, número de visualizações ou citações e ano de lançamento.
- O primeiro paper a ser lido foi um survey de Multimodal Models. Apesar de não ter finalizado a leitura, foi possível obter alguns insights:
    - MLLMs surgiram com modelos como GPT-4V, unindo texto e imagens e trazendo novas capacidades (gerar código a partir de imagens, interpretar memes, resolver problemas sem OCR);
    - São vistos como um passo em direção à AGI;
    - A arquitetura típica combina encoder de modalidade, conector e LLM pré-treinado, mas também existem modelos que unificam modalidades em um único encoder, como o ImageBind.
- Documento detalhado com a lista de conteúdos selecionados, bem como a forma que as

pesquisas foram feitas e lista de prioridades pode ser acessado pelo link:
Residência - Semana 1 - Documento Interno

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

- Concluir a leitura detalhada da survey “A survey on multimodal large language models”.
- Assistir aulas prioritárias da lista (como a Lecture 8 – MIT e Stanford CS25).
- Produzir anotações organizadas e resumos sobre os conteúdos lidos.
- Consolidar um entendimento claro das bases e fundamentos dos Multimodal Models, para poder avançar no campo das arquiteturas atuais e aplicações.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

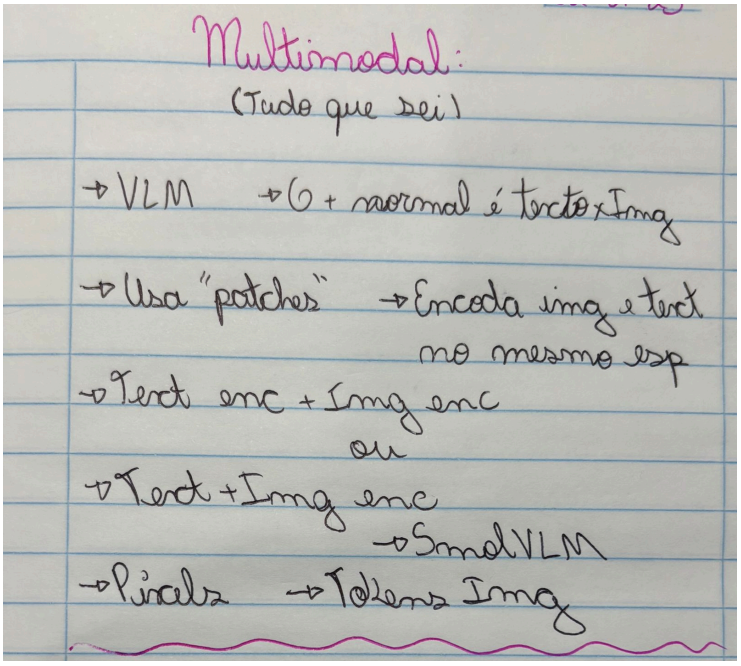
CEDRIC LUIZ DE CARVALHO: Go!

[Residência - Semana 1 - Documento Interno citado no Termo de Aceite de Entrega de 03 de setembro]

Semana 1

Primeiro contato com o tema:

Nessa primeira semana, meu foco é inteiramente entender melhor os fundamentos do tema. Antes de consumir qualquer conteúdo sobre ele, apliquei a técnica de pegar um papel, colocar o tema como título (Multimodal) e escrever tudo que eu sei ou já ouvi falar sobre o tema em termos simples e básicos. O resultado foi esse:



De acordo com a foto, é possível ver que meu conhecimento do tema é bem limitado. Muitas coisas me lembram da minha área de atuação (processamento de linguagem natural), mas ao mesmo tempo tenho algumas (muitas) dúvidas:

- O que eles usam para codificar imagem?
- Como imagens e textos se comunicam?
- Porque não temos muitas IAs multimodais de texto e áudio, ou imagem e áudio?
- Quais são os melhores modelos de IA Multimodal atualmente?

- Como posso aplicar a IA Multimodal da melhor forma?
- Qual foi a primeira arquitetura de IA Multimodal?
- Qual a história dos modelos Multimodais?
- Os modelos Multimodais foram criados com qual propósito?
- Quais possíveis sub-áreas dentro de Multimodal?

Por mais que eu gostaria de ter a resposta para todas as perguntas hoje ainda, sei que com o passar das semanas essas respostas vão vir naturalmente, e esse é meu objetivo no começo do processo: **entender melhor o que é, qual a história, quais são as aplicações, como foi o passado, como é o presente e como possivelmente será o futuro das Inteligências Artificiais Multimodais.**

Pesquisa de Conteúdos Úteis:

Para entender melhor como surgiu e o que temos de Multimodal hoje em dia, decidi pesquisar papers academicamente relevantes sobre o tema. Para isso, pesquisei no Semantic Scholar e no Google Scholar, com a query “survey multimodal” e ordenação por relevância.

Obtive vários resultados, e selecionei os mais interessantes para minha pesquisa de acordo com o abstract:

- Efficient Multimodal Large Language Models: A Survey (2024 - Citado por 87);
- Explainable and interpretable multimodal large language models: A comprehensive survey (2024 - Citado por 41);
- A Survey of Multimodal Large Language Model from A Data-centric Perspective (2024 - Citado por 59);
- The Revolution of Multimodal Large Language Models: A Survey (2024 - Citado por 117);
- A survey on multimodal large language models (2024 - Citado por 2010);
- A comprehensive survey and guide to multimodal large language models in vision-language tasks (2024 - Citado por 20);
- Mm-llms: Recent advances in multimodal large language models (2024 - Citado por 436).

Além disso, para entender posteriormente o cenário atual das arquiteturas e aplicações do tema atualmente, pesquisei “Multimodal Models” no Semantic Scholar e filtrei para resultados no ano de 2025:

- Llava-mini: Efficient image and video large multimodal models with one vision token (2025 - Citado por 54);
- BLIP3-o: A Family of Fully Open Unified Multimodal Models-Architecture, Training and Dataset (2025 - Citado por 18);

- SmolVLM: Redefining small and efficient multimodal models (2025 - Citado por 36);
- Show-o2: Improved Native Unified Multimodal Models (2025 - Citado por 8).
- IMAGEBIND: One Embedding Space To Bind Them All (2023 - Citado por 1295).

Para complementar os estudos e não me basear apenas na leitura de papers, decidi buscar também alguns vídeos no YouTube sobre o tema. Procurei com a palavra chave “multimodal models class”, para que a pesquisa me apresentasse AULAS sobre o tema, e não apenas vídeos explicativos. Com isso, achei alguns conteúdos interessantes:

- Multimodal Machine Learning (playlist) (2024)
- Multimodal Learning at CVPR 2022 (playlist) (2022)
- Stanford CS224N NLP with Deep Learning | 2023 | Lecture 16 - Multimodal Deep Learning (video) (2023)
- CS 198-126: Lecture 22 - Multimodal Learning (video) (2023)
- Stanford CS25: V4 I From Large Language Models to Large Multimodal Models (video) (2024)
- LLM Chronicles #6.3: Multi-Modal LLMs for Image, Sound and Video (video) (2024)
- Stanford CS25: V5 I Multimodal World Models for Drug Discovery, Eshed Margalit of Noetik.ai (video) (2025)
- Lecture 8 – Large Multimodal Models (MIT How to AI Almost Anything, Spring 2025) (video) (2025)

Com esses conteúdos, tenho um caminho a seguir. São muitos papers e muitos vídeos, todos longos e com muito conteúdo. Por isso, não é possível ler todos ou acessar todos os conteúdos. Dessa forma, eu criei uma lista de prioridades para decidir qual conteúdo priorizar. Essa lista pode mudar conforme eu acesso o conteúdo, ou vejo algum outro conteúdo mais interessante. É um caminho que não está totalmente definido, podendo haver mudanças no trajeto, mas o destino é o mesmo: entender a história e ter a base de Multimodal Models!

Lista de Prioridades:

Como dito anteriormente, eu reuni muitos conteúdos. Para otimizar o tempo e não passar mais de uma semana entendendo os fundamentos, decidi criar uma lista de prioridades e colocar os conteúdos em ordem para serem consumidos. Eu fiz essa lista de acordo com o assunto abordado, número de visualizações ou citações e ano de lançamento.

É importante salientar que essa lista pode mudar, e que provavelmente não será possível ler todos os conteúdos à tempo. A lista de prioridade ficou a seguinte:

1. A survey on multimodal large language models (2024 - Citado por 2010)
2. Lecture 8 – Large Multimodal Models (MIT How to AI Almost Anything, Spring 2025)
3. Mm-llms: Recent advances in multimodal large language models (2024 - Citado por 436)
4. Stanford CS25: V4 I From Large Language Models to Large Multimodal Models (2024)
5. Stanford CS224N NLP with Deep Learning | Lecture 16 - Multimodal Deep Learning (2023)
6. Multimodal Machine Learning (playlist) (2024)
7. The Revolution of Multimodal Large Language Models: A Survey (2024 - Citado por 117)
8. LLM Chronicles #6.3: Multi-Modal LLMs for Image, Sound and Video (2024)
9. Multimodal Learning at CVPR 2022 (playlist) (2022)
10. ImageBind: One Embedding Space To Bind Them All (2023 - Citado por 1295)
11. CS 198-126: Lecture 22 - Multimodal Learning (2023)
12. A Survey of Multimodal Large Language Model from A Data-centric Perspective (2024 - Citado por 59)
13. Efficient Multimodal Large Language Models: A Survey (2024 - Citado por 87)
14. Explainable and interpretable multimodal large language models: A comprehensive survey (2024 - Citado por 41)
15. A comprehensive survey and guide to multimodal large language models in vision-language tasks (2024 - Citado por 20)
16. Llava-mini: Efficient image and video large multimodal models with one vision token (2025 - Citado por 54)
17. SmolVLM: Redefining small and efficient multimodal models (2025 - Citado por 36)
18. BLIP3-o: A Family of Fully Open Unified Multimodal Models-Architecture, Training and Dataset (2025 - Citado por 18)
19. Show-o2: Improved Native Unified Multimodal Models (2025 - Citado por 8)
20. Stanford CS25: V5 I Multimodal World Models for Drug Discovery, Eshed Margalit of Noetik.ai (video) (2025)

Com o conteúdo definido e prioridades decididas, chegou a hora de começar a consumir esses conteúdos, gerando resumos e anotações com as descobertas. A começar pelo primeiro paper: “A survey on multimodal large language models”.

A Survey on Multimodal Large Language Models

Ainda não foi possível ler a survey completamente, contudo, já consegui ter alguns insights:

- MLLMs surgiram com modelos como o **GPT-4V**, capazes de entender e raciocinar sobre texto e imagens.
- Trazem novas capacidades, como escrever código a partir de imagens, interpretar memes e resolver problemas matemáticos sem OCR.
- São considerados um passo promissor em direção à inteligência artificial geral (AGI).
- Estrutura típica envolve:
  - **Encoder de modalidade** (ex.: CLIP, EVA-CLIP, ImageBind).
  - **LLM pré-treinado** (LLaMA, Vicuna, Qwen, etc.).
  - **Interface de modalidade (conector)** para alinhar representações visuais/áudio com texto (via projeção, queries ou fusão).
- Alguns modelos também incluem **geradores** para saída em múltiplas modalidades (ex.: imagens e vídeos).

No survey eles explicam que os MLLMs mais comuns seguem a arquitetura modular (encoder de imagem + conector + LLM). Mas também existem os modelos que buscam unificar diretamente visão e linguagem em um mesmo encoder/representação. Dessa maneira, foi incluído o paper do ImageBind na minha lista, pois ele cria um espaço comum multimodal (imagem, texto, áudio, etc.), de onde as representações são extraídas de forma conjunta.

Próximos Passos

- Continuar leitura do paper “A survey on multimodal large language models”;
- Assistir as aulas para entender melhor o conteúdo;
- Fazer anotações com pontos importantes e marcações inteligentes;
- Ao fim da Semana 2, ter um conhecimento concreto do que são Multimodal Models e as arquiteturas mais usadas hoje em dia.

APÊNDICE 2

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: **11 de set. de 2025**

Participantes da Entrega **[matriculados em Residência em IA]**:

Fernanda Bufon Farber

Entrega: **[descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]**

- Concluí a leitura do survey "A Survey on Multimodal Large Language Models", destacando pontos importantes ao longo do texto para referência futura.
- Iniciei o acompanhamento das aulas prioritárias da lista, começando pela Lecture 8 do MIT, que trouxe uma visão inicial sobre como os modelos multimodais são estruturados em termos de fundamentos e arquiteturas.
- Um dos principais aprendizados foi que Multimodal Models e MLLMs não são exatamente a mesma coisa:
- Multimodal Models:
    - Abrangem qualquer modelo de IA que integra e processa múltiplas modalidades de dados, como texto, imagens, áudio ou vídeo, buscando representações conjuntas para raciocínio mais completo.
- MLLMs (Multimodal Large Language Models):
    - Representam um subconjunto dos modelos multimodais, em que um LLM pré-treinado em linguagem é estendido com encoders e conectores para outras modalidades, permitindo que o modelo interprete e gere respostas a partir da combinação de texto e outros tipos de dados.
- Essa distinção me ajudou a entender que multimodalidade é um campo muito mais amplo, enquanto os MLLMs são uma vertente específica que me interessaram bastante, e desejo entender mais e seguir nesse caminho.

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

- Produzir anotações detalhadas e resumos organizados a partir da leitura do survey no caderno;
- Fazer uma lista de papers sobre MLLMs que foram citadas no survey;
- Ler outro survey para ter mais informações e outra visão sobre Multimodal Models;
- Avançar na lista de aulas prioritárias, incluindo a *Stanford CS25*;
- Entender melhor as bases e fundamentos dos modelos multimodais;

Observação: **[caso precise fazer alguma observação, de qualquer "natureza"]**

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

APÊNDICE 3

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 18 de set. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

Fernanda Bufon Farber

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Aprofundei a compreensão sobre o funcionamento dos MLLMs.
- Procurei e selecionei papers atuais para analisar as arquiteturas multimodais mais recentes e mais usadas.
- Identifiquei um possível caminho de desenvolvimento: a interação com o usuário pode ser mais empática se o LLM tiver capacidade de entender emoções.
    - Apenas o texto não é suficiente para capturar sentimentos; por isso, é interessante que o modelo consiga interpretar expressões visuais.
- Fiz uma lista com os modelos mais usados atualmente (open-source ou não) no contexto multimodal.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Usar o NotebookLM para consolidar resumos e entender melhor cada paper e arquitetura.
- Realizar um primeiro contato com modelos multimodais (abertos e fechados) para testar e entender o funcionamento prático.
- Explorar a possibilidade de que modelos multimodais interpretem expressões faciais e emoções humanas.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

APÊNDICE 4

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 25 de set. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

Fernanda Bufon Farber

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Aprofundei o estudo das arquiteturas multimodais utilizando o NotebookLM, que serviu como ferramenta para consolidar os papers e gerar conteúdos didáticos (vídeos, podcasts e mapas mentais) a partir dos papers analisados. O Notebook pode ser acessado aqui.
- Avancei da pesquisa teórica para a organização prática de testes, construindo uma planilha estruturada com diferentes modelos multimodais disponíveis, detalhando como acessá-los (API, código, interface). A planilha está no Notion e pode ser acessada aqui.
- Identifiquei e analisei de forma inicial o paper R1-Omni e o paper HumanOmni, que propõe uma abordagem multimodal para entendimento de vídeos com humanos e detecção de emoções, apontando que há um esforço por meio das arquiteturas de atacar esse tema.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Realizar leitura aprofundada dos papers Omni-R1 e HumanOmni, compreendendo suas arquiteturas, metodologias e resultados.
- Executar testes práticos com os modelos multimodais já listados, registrando observações sobre usabilidade e desempenho.
- Iniciar um levantamento de trabalhos que aplicam modelos multimodais em detecção de emoções, buscando identificar tendências e lacunas de pesquisa na área.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

[Documento modelos.xlsx citado no Termo de Aceite de Entrega de 25 de Setembro. A planilha está em formato Google Sheets para melhor compreensão.]

Name	Input Price	Link	Observations
Realtime API (OpenAI)	$5.00	https://platform.openai.com/docs/guides/realtime, https://platform.openai.com/docs/models/gpt-realtime	TEM QUE ESTUDAR MAIS PRA VER QUAL MODELO USA.
LLaVA (versão padrão / 1.5)			Facilmente adaptável mas desempenho menor em tarefas visuais complexas comparado a modelos proprietários
LLaVA-NeXT			Mais robusto que versões anteriores mas ainda em evolução
Qwen-VL / Qwen2-VL / Qwen2.5-VL			Inclui suporte para OCR e bounding boxes
NExT-GPT			Modelo ambicioso mas ainda experimental
Macaw-LLM			Mais protótipo de pesquisa do que ferramenta pronta
VARGPT			Modelo de pesquisa emergente com possíveis limitações de estabilidade
LLaVA-Mini			Alta eficiência mas perde detalhes visuais finos
LLaVA-UHD			Especialmente útil para imagens grandes e formatos não convencionais. NÃO TENHO MEMÓRIA
SmolVLM			É um modelo pequeno
R1-Omni		https://r1-omni.com/	Feito pra entender sentimentos. R1-Omni is an AI emotion recognition model that integrates Reinforcement Learning with Verifiable Reward (RLVR) into an Omni-multimodal framework. Open-source.
GPT-5		https://platform.openai.com/docs/models/gpt-5	Meio caro…

GPT-5-Mini		https://platform.openai.com/docs/models/gpt-5-mini	Menor e mais rápido
GPT-5-Nano		https://platform.openai.com/docs/models/gpt-5-nano	Bem mais barato, mas mto pequeno, pode não ser tão bom.
GPT-4.1		https://platform.openai.com/docs/models/gpt-4.1	
GPT o3-pro	$20.00	https://platform.openai.com/docs/models/o3-pro	Usa reasoning avançado
GPT o4-mini	$1.10	https://platform.openai.com/docs/models/o4-mini	
GPT-4o mini	$0.15	https://platform.openai.com/docs/models/gpt-4o-mini	
GPT 4.1-Mini	$0.40	https://platform.openai.com/docs/models/gpt-4.1-mini	
GPT-4.1 Nano	$0.10	https://platform.openai.com/docs/models/gpt-4.1-nano	
GPT-4o mini Realtime	$0.60	https://platform.openai.com/docs/models/gpt-4o-mini-realtime-preview	
GPT-4o Realtime	$5.00	https://platform.openai.com/docs/models/gpt-4o-realtime-preview	
Gemini			

APÊNDICE 5

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 2 de out. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

Fernanda Bufon Farber

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Realizei a leitura do paper Omni-R1, aprofundando a compreensão de sua arquitetura, metodologia e resultados.
- Executei testes práticos com três modelos multimodais já listados anteriormente (todos da família GPT), entendendo melhor como utilizar e seu desempenho.
- Realizei uma pesquisa exploratória com papers que fazem testes nos modelos multimodais, para identificar quais MLLMs são mais adequados para o meu contexto de uso, avaliando prós e contras de diferentes opções.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Continuar os testes práticos com modelos multimodais.
- Identificar e analisar em detalhe os Top 3 modelos multimodais para uso prático, considerando critérios como desempenho, acessibilidade e documentação.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

APÊNDICE 6

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 9 de out. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

Fernanda Bufon Farber

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Escolhi os modelos que vou priorizar para os testes práticos.
- Sistematizei o processo de testes, definindo dois datasets, a quantidade de dados por categoria e critérios de avaliação para manter consistência entre os experimentos. O documento descrevendo o processo pode ser encontrado no documento Experimento.docx.
- Criei um repositório no github e uma organização no HugginFace para versionar os dados e códigos usados.
- Organizei e subi os dados no Hugging Face, facilitando o acesso durante os testes.
- Reuni trabalhos que analisam as capacidades multimodais dos MLLMs, mas optei por também realizar experimentos práticos para compreender como utilizar cada modelo na prática, explorando formas de uso e de integração.
- Avancei na familiarização com os modelos selecionados, aprendendo quais ferramentas (frameworks) usar e como acessar os modelos.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Executar os testes práticos e sistematizados com os MLLMs selecionados, com foco nos modelos acessíveis via API (sem modelos locais nesse momento).
- Ler e sintetizar papers reunidos que comparam a qualidade e capacidades gerais dos LLMs multimodais de forma geral.
- Iniciar uma pesquisa sobre análise de sentimento em tempo real com LLMs, verificando trabalhos semelhantes e bibliotecas disponíveis para detecção de sentimentos em tempo real.
- Estruturar os resultados obtidos no teste em formato comparativo.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

[Experimento.docx citado no Termo de Aceite de Entrega de 9 de Outubro.]

Ideia

- Usar dois tipos de dados para avaliar **modelos multimodais de reconhecimento de emoções**. Um dos datasets é mais difícil de classificar do que o outro.
  O objetivo é verificar se o modelo classifica corretamente a emoção (sim ou não).

Configuração

Datasets:

Mais fácil (FEMO):

As expressões faciais são mais claras e a qualidade das imagens é bem melhor. Porém, temos menos dados.

Facial Emotion Recognition Dataset

Mais difícil (FEXP):

As imagens são menores, em preto e branco, e as expressões faciais são mais sutis. O dataset é maior.

Face expression recognition dataset

Informações dos Dados:

Emoções:

- Neutro
- Surpresa
- Triste
- Feliz
- Medo
- Nojo

- Raiva
- Desprezo (presente apenas no primeiro dataset)

Número de Amostras:

Vamos usar 30 amostras por emoção por questão de orçamento:

- 15 do FEMO;
- 15 do FEXP;
- Obs.: Desprezo terá apenas 15 amostras.

Usei um código em Python para selecionar aleatoriamente 15 amostras de cada emoção em cada dataset. Salvei 2 datasets diferentes para acessar as pontuações separadamente.

Modelos:

Vamos avaliar os seguintes modelos multimodais (detalhes e links estão organizados no Notion):

- R1-Omni
- Realtime API (OpenAI)
- GPT-5
- GPT-5-Mini
- GPT-5-Nano
- GPT-4.1
- GPT-4.1-Mini
- GPT-4.1 Nano
- GPT o3-pro
- GPT o4-mini
- GPT-4o mini
- GPT-4o mini Realtime
- GPT-4o Realtime
- Gemini
- SmolVLM
- LLaVA-UHD
- LLaVA-Mini
- VARGPT
- Macaw-LLM
- NExT-GPT

Pipeline:

1. Escolher o modelo;
2. Passar as imagens para o modelo com um prompt padrão e simples (a ser definido);
3. Obter as classificações;
4. Calcular as pontuações;
5. Salvar as estatísticas.

APÊNDICE 7

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 16 de out. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

Fernanda Bufon Farber

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Li o survey que compara a capacidade dos MLLMs, procurando entender a capacidade de cada um e quais são as métricas usadas. Além disso, o paper dá um bom background dos avanços na área.
- Não foram feitos testes práticos pois identifiquei a necessidade de ampliar a base de dados, uma vez que um conjunto de apenas 100 amostras seria insuficiente para conclusões consistentes. Por isso, estou buscando em papers práticos algum benchmark ou framework de avaliação.
- Reuni e filtrei artigos relacionados à aplicação de MLLMs em detecção de sentimentos, realizando leitura dos abstracts e separando os trabalhos mais relevantes para estudo detalhado.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Ler os papers selecionados sobre detecção de sentimentos com MLLMs, buscando entender os métodos e tecnologias empregados em cada um, bem como a avaliação.
- Escolher um paper viável para replicação experimental, priorizando aqueles que utilizam modelos open-source e menores.
- Continuar o preenchimento da planilha com informações sobre modelos multimodais, incluindo nome, parâmetros e capacidades. Esse material poderá ser útil na etapa de experimentação.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

APÊNDICE 8

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 23 de out. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

Fernanda Bufon Farber

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Li os papers selecionados, focando em um específico que cria um framework para melhorar a detecção de sentimento sem ter que realizar o fine-tuning dos modelos (Catch Your Emotion: Sharpening Emotion Perception in Multimodal Large Language Models).
- Sobre o paper lido, tive algumas observações:
  - O método deles trouxe resultados melhores que o baseline (zero-shot), mas a melhora foi muito pequena e ainda assim insatisfatória.
  - Os modelos que usaram são open-source, mas é preciso máquinas potentes para rodá-los.
  - Não senti confiança nos métodos que eles usaram.
- Com isso, decidi procurar outro trabalho com melhores resultados, mesmo que use outros métodos.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Escolher outro paper para replicar o que fizeram.
- Fazer a replicação de um experimento.
- Fazer também, paralelamente, uma v0 simples contendo uma pipeline básica de detecção de sentimento (imagem -> API de um MLLM pago -> resultado) para "por a mão na massa" e entender se é viável usá-los.
- Entender se há algum modelo open-source que possa ser usado.
- Fazer testes com a V0.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

APÊNDICE 9

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 6 de nov. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

Fernanda Bufon Farber

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Fiz uma busca extensiva para achar um paper bom para replicar. A tabela de análise de paper pode ser encontrada no documento Analise_Papers.xlsx
- Fiz uma v0 simples contendo uma pipeline básica de detecção de sentimento.
- Testei a versão 0 usando o modelo Gemini. Não foi possível implementar um modelo open por enquanto.
- Encontrei um paper bom para replicar (Visual Prompting in LLMs for Enhancing Emotion Recognition).
- Comecei a organizar como serão feitos os experimentos. Organizei o experimento no documento Experimentos_v2.docx.
- Comecei a seguir o pipeline do paper em proporções menores.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Continuar a replicação.
- Escolher modelos multimodal open-source para fazer o teste.
- Detalhar os resultados.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

[Analise_Papers.xlsx citado no Termo de Aceite de Entrega de 6 de Novembro.]

Name	Can Replicate?	Can evaluate?	Computing	Method	Priority	Reading Order
EmoGist: Efficient In-Context Learning for Visual Emotion Understanding	Yes	No	Not too much, but a lot	Zero-shot	Medium	1
UniMEEC: Towards Unified Multimodal Emotion Recognition and	No	No	-	Training	Low	7
Multimodal Emotion Captioning Using Large Language Model with Prompt Engineering	Yes	No	-	Zero-shot	Medium	2
Visual and Textual Prompts in VLLMs for Enhancing Emotion	Yes	No	Can use API	Zero-shot	High	4
BeMERC: Behavior-Aware MLLM-based Framework for Multimodal	No	No	A lot.	Training	Low	8
MELT: Towards Automated Multimodal Emotion Data Annotation by	No	No	-	Data Creation	Low	14

CUSTOMIZING VISUAL EMOTION EVALUATION FOR MLLMS: AN OPEN-VOCABULARY, MULTIFACETED, AND SCALABLE APPROACH	No	No	-	Data Creation	Low	10
Emotion Knowledge Enhancement for Vision Large Language Models: A Self-Verification Approach for High-Quality Emotion	No	No	-	Data Creation	Low	11
GPT-4V with Emotion: A Zero-shot Benchmark for Generalized Emotion Recognition	No	No	-	Data Creation	Low	5
Emotion-Qwen: A Unified Framework for Emotion and Vision Understanding	No	No	3 NVIDIA A800 80GB GPUs	Training	Low	3
OV-MER: Towards Open-Vocabulary Multimodal Emotion Recognition	No	No	EXPENSIVE	Data Creation	Low	13
Explainable Multimodal Emotion Recognition	No	No	-	Data Creation, Zero-shot	Low	9

Multimodal Video Emotion Recognition with Reliable Reasoning Priors	No	No	YES	Training	Low	15
FEALLM: Advancing Facial Emotion Analysis in Multimodal Large Language Models with Emotional Synergy and Reasoning	No	No	Expensive	Data Creation, Training	Low	16
More Is Better: A MoE-Based Emotion Recognition Framework with Human Preference Alignment	No	No	-		Low	17
EmoVerse: Exploring Multimodal Large Language Models for Sentiment and Emotion Understanding	No	No	Will take a lot of hours.	Data Creation, Training	Low	18
Face-LLaVA: Facial Expression and Attribute Understanding through Instruction Tuning	No	Yes	8 H100 to replicate, 4090 INFERENCE	Data Creation, Training	High	19

Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning	No	Yes	4 A100 to replicate, RTX 4090 INFERENCE	Training	High	20
Enhancing Emotion Reasoning for Image Multi-Emotion Prediction	No	No	-	Fine-tuning	Low	21
Facial Emotion Detection Research Based on an Improved Multi-modal LLM	No	No	2x A100 to replicate	Fine-tuning	Low	22
Emotion Recognition from Videos Using Multimodal Large	Yes	No	A6000, but can try to replicate in a small setting.	Fine-tuning, Zero-shot	High	23

[Experimentos_v2.docx citado no Termo de Aceite de Entrega de 6 de Novembro.]

Definição:

Estou testando o método proposto no artigo “Visual and Textual Prompts in VLLMs for Enhancing Emotion Recognition”.

Pretendo reproduzir o paper de forma mais fiel possível ao original, mas, por questões de tempo, vou concentrar meus esforços nas partes mais relevantes para o meu contexto, especialmente aquelas que possam ser aplicadas à detecção de emoções por webcams de computador.

Obs:

É difícil encontrar um dataset que utilize postura corporal para o reconhecimento de emoções — a maioria dos conjuntos de dados disponíveis foca apenas em expressões

faciais. Buscar, coletar e anotar novos dados que incluam informações posturais, seja por meio de anotação manual ou geração sintética.

Meu objetivo é verificar, em pequena escala, se o uso de prompts visuais e textuais (SoVTP) realmente melhora a interpretação emocional de modelos multimodais (VLLMs), comparado a abordagens puramente textuais.

Arquitetura do pipeline

Obs: salvar imagens após cada iteração.

→ Box Detection:

- Usa RetinaFace.
- Bom pra saber no que devemos focar.
- Faz uma bounding box.
- ~~Acho que o código está disponível.~~
- **Dá pra usar o Py-feat para tudo.**

→ Muscle Movement Detection and Analysis:

- Landmarks.
- Action Units.
- Feito com o uso do Py-Feat.

→ Colocar informações na imagem (igual no paper):

- Colocar apenas a bounding box e os pontos.
- Não faz sentido colocar outras coisas.
- Testar sem também.

→ Escrever informações que o Py-Feat passou.

→ OpenAI

→ Passar imagens para o VLLM de acordo com os prompt:

Prompt 1 - Linguagem Corporal

- **Objetivo**: capturar expressões corporais e gestos que complementam as expressões faciais.

- **O modelo analisa:**

    - postura (curvado, ereto, relaxado, rígido);

  - direção do corpo e das mãos;
  - gestos que indicam tensão, felicidade ou nervosismo.
- Exemplo de prompt:
  "What is the body language of the person?"

Prompt 4 - Movimentos Musculares / Action Units (AUs)

Obs: Ver melhor como isso funciona, como o DeepFace retorna os AUs.

- **Objetivo:** identificar quais músculos da face estão ativos e associá-los às emoções básicas.

- **O modelo analisa:**

  - *Action Units* detectados (ex.: AU06 – cheek raise, AU12 – lip corner puller, AU25 – lip part);
  - intensidade de cada AU;
  - combinação de AUs que indica emoções específicas (felicidade, tristeza, raiva etc.).
- **Exemplo de prompt:**
  "Which facial action units (AUs) are active for the person, and what do they suggest about the emotion?"

Prompt 5 - Integração e Autocorreção (Self-Correction Stage)

- **Objetivo:** integrar todas as informações anteriores — contexto, corpo, emoções de outros e AUs — para concluir a emoção final com base em evidências.

- **O modelo analisa:**

  - coerência entre o ambiente, o corpo e as expressões faciais;
  - reforço ou contradição entre as pistas observadas;
  - emoção mais provável considerando todos os fatores.
- **Exemplo de prompt:**
  "Using the context, body language, others' emotions, and facial action units, what is the emotion of the person?"

→ Inferência no VLLM (ChatGPT, Gemini, depois modelo open)

→ Avaliação (Accuracy e F1)

- Podemos ver outras métricas no futuro.(SoVTP preserva o contexto da cena e adiciona camadas visuais; o prompting em etapas integra contexto, corpo, AUs e emoções de terceiros antes da resposta final.)

Passos detalhados

1) Implementar Código

- Fazer código e usar imagens FER para testar se está correto.
- Há um código no github que pode ser útil.
- Implementar todo o pipeline.

2) Testar Pipeline

- Usar dataset pequeno para testar e validar o pipeline de tratamento dos dados e prompts.

3) Escolher Dataset fFinal

- Procurar um dataset que seja mais parecido com nosso caso;
- Visualizar qual é a melhor forma de fazer um dataset caso não seja possível achar um dataset interessante.
- Podemos usar um grande conjunto de dados com labels + algumas amostras de webcam (as vezes 15 de cada emoção, anotados).

4) Organizar Experimento

- Definir e implementar os experimentos:
    - Definir quais modelos vamos usar
        - GPT
        - Gemini
        - QwenVL pequeno
    - Definir exatamente quais cenários vamos ter
        - Only Py-Feat.
        - Only DeepFace
        - VLM imagem normal (prompt básico + imagem)
        - VLM com Py-Feat na imagem
        - VLM com Py-Feat em texto
        - VLM com Prompts
        - ETC

APÊNDICE 10

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 13 de nov. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

Fernanda Bufon Farber

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

- Continuação da replicação do paper.
- Obtive resultados muito inferiores ao esperado, mas estou buscando usar o mesmo modelo que usaram.
- Tive algumas ideias que podem aumentar o desempenho do modelo, sem ser o método do paper.
- Os resultados preliminares podem ser analisados no arquivo Resultados.docx

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

- Finalizar replicação.
- Reportar resultados.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

[Resultados.docx citado no Termo de Aceite de Entrega de 13 de Novembro.]

Resultados

Neste documento, apresentamos uma análise abrangente do desempenho dos modelos multimodais de linguagem (MLLMs) avaliados na tarefa de detecção de emoções a partir de imagens faciais. Foram examinados três modelos — *gpt5nano*, *qwen_instruct* e *qwen* — em combinação com quatro métodos distintos de prompting: **only_image**, **only_image_simple**, **sovtp** e **sovtp_sameimg**.

Os dois primeiros métodos representam abordagens de prompting direto: em **only_image**, a imagem é acompanhada por um prompt composto, estruturado para guiar o modelo por diferentes aspectos visuais relevantes; em **only_image_simple**, a imagem é acompanhada apenas por uma instrução direta e concisa ("Qual emoção a pessoa está sentindo?"). Os métodos **sovtp** e **sovtp_sameimg** correspondem à reprodução do método apresentado por Wang et al. (2025), em que visual e textual prompts são empregados de forma sequencial. No método original (**sovtp**), cada etapa do pipeline recebe uma imagem modificada com realce visual da característica correspondente; já em **sovtp_sameimg**, todas as etapas utilizam a mesma imagem original, sem variações visuais, permitindo avaliar o quanto o componente multimodal depende da construção progressiva de sinais visuais.

Resultados Globais

O primeiro conjunto de análises buscou comparar o desempenho agregado dos modelos e métodos. A Tabela a seguir apresenta acurácia, F1 macro, recall macro e precision macro para cada combinação modelo × método, permitindo uma visão integrada da performance global.

Model	Method	Accuracy	F1_macro	Recall	Precision
GPT-5 Nano	Only Image	0.29	0.20	0.26	0.19
	Only Image (simple prompt)	0.32	0.28	0.32	0.46
	SoVTP	0.32	0.28	0.32	0.29
	SoVTP (same image)	0.27	0.21	0.28	0.20
Qwen-VL 8B	Only Image (simple prompt)	0.29	0.28	0.30	0.23
	SoVTP	0.19	0.16	0.19	0.27
Qwen-VL 8B Instruct	Only Image	0.32	0.24	0.28	0.29

	Only Image (simple prompt)	0.34	0.28	0.34	0.29
	SoVTP	0.29	0.24	0.29	0.36
	SoVTP (same image)	0.29	0.24	0.29	0.23

Tabela 1 - Desempenho global dos modelos GPT-5 Nano, Qwen-VL 8B e Qwen-VL 8B Instruct nos diferentes métodos (Only Image, Only Image – simple prompt, SoVTP e SoVTP – same image), medido por acurácia, F1 macro, recall macro e precisão.

A observação geral desses resultados mostra que o método **only_image_simple** apresenta, de forma consistente, os melhores valores de F1 macro entre os três modelos avaliados. Isso sugere que, no contexto testado, prompts simples e diretos podem ser mais eficazes do que prompts compostos ou pipelines multimodais mais elaboradas. O modelo **qwen_instruct** demonstra desempenho superior de maneira geral, tanto em F1 macro quanto em acurácia, seguido pelo **gpt5nano** e, por último, pelo **qwen**, que obtém métricas mais modestas.

Em contraste, os métodos **sovtp** e **sovtp_sameimg**, apesar de inspirados em uma pipeline mais complexa e originalmente eficaz em configurações específicas (Wang et al., 2025), não apresentaram vantagens sistemáticas neste conjunto experimental. Embora **sovtp** demonstre ganhos pontuais em emoções negativas como tristeza ou surpresa, não houve impacto positivo na métrica global.

Desempenho por Classe

Para compreender melhor a sensibilidade dos modelos diante de cada emoção, realizamos uma análise detalhada por classe. As tabelas seguintes apresentam o F1-score por emoção (anger, disgust, fear, happy, neutral, sad, surprise) para cada método dentro de cada modelo.

Model	Method	anger	disgust	fear	happy	neutral	sad	surprise
GPT-5 Nano	Only Image	0.2	0.0	0.0	0.68	0.35	0.25	0.15
	Only Image (simple prompt)	0.33	0.09	0.08	0.58	0.32	0.33	0.17
	SoVTP	0.28	0.08	0.0	0.48	0.41	0.39	0.22
	SoVTP (same image)	0.27	0.0	0.0833	0.62	0.31	0.17	0.0
Qwen-VL 8B	Only Image (simple prompt)	0.2	0.0	0.0	0.62	0.36	0.29	0.17
	SoVTP	0.07	0.0	0.1818	0.26	0.08	0.24	0.26
Qwen-VL	Only Image	0.08	0.0	0.25	0.61	0.30	0.36	0.32

8B Instruct	Only Image (simple prompt)	0.13	0.0	0.087	0.58	0.42	0.42	0.30
	SoVTP	0.08	0.09	0.0	0.52	0.25	0.46	0.22
	SoVTP (same image)	0.0	0.0	0.08	0.60	0.29	0.46	0.20

Tabela 2 – F1-score por classe de emoção (anger, disgust, fear, happy, neutral, sad e surprise) para cada combinação de modelo (GPT-5 Nano, Qwen-VL 8B e Qwen-VL 8B Instruct) e método de prompting.

A análise dessas tabelas evidencia padrões consistentes entre os modelos. A classe **happy** apresenta o maior F1 em todos os casos, com valores frequentemente superiores a 0.55, indicando que sorrisos e expressões de alegria são mais facilmente capturados pelos MLLMs. As classes **neutral** e **sad** aparecem em seguida, com desempenho intermediário.

Por outro lado, as classes **disgust** e **fear** apresentam desempenho extremamente baixo, muitas vezes com F1 próximo de zero. Esse padrão sugere que tanto os dados quanto os modelos têm dificuldade em representar expressões mais discretas ou menos evidentes, possivelmente devido ao número reduzido de amostras ou à baixa saliência visual dessas emoções. Essa limitação aparece de maneira consistente entre todos os métodos e modelos.

O método **only_image_simple** também se destaca no nível por classe, sendo o mais equilibrado e robusto, especialmente para anger, neutral e sad. Já o método **sovtp** apresenta desempenho levemente superior em algumas emoções negativas — como surpresa e tristeza — mas esse ganho não se traduz em um aumento global de desempenho. O método **sovtp_sameimg**, por sua vez, apresenta quedas significativas em algumas classes, sugerindo que o pipeline original depende fortemente das variações visuais usadas em cada etapa.

Dificuldade Relativa Entre Emoções

Para sintetizar a dificuldade relativa de cada classe, calculamos o F1-score médio por classe ao longo de todos os modelos e métodos. O gráfico correspondente facilita a visualização da hierarquia de dificuldade entre as emoções.

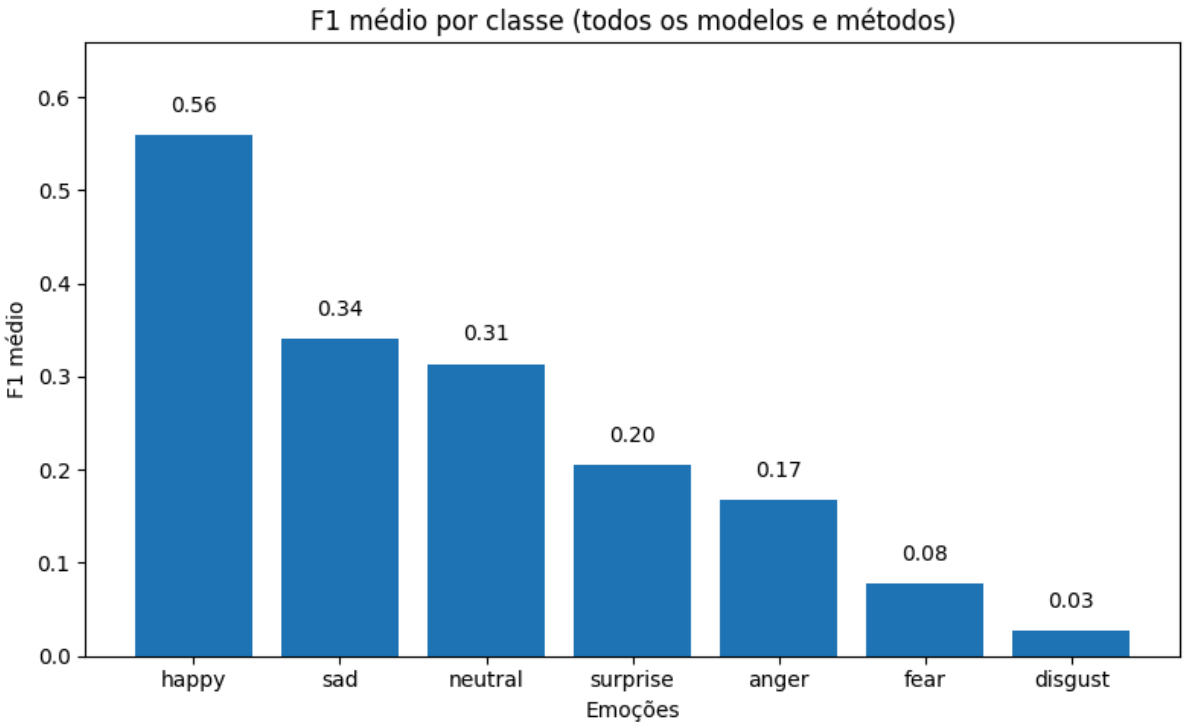


Figura 1 – F1-score médio por classe de emoção (anger, disgust, fear, happy, neutral, sad e surprise), considerando conjuntamente todos os modelos e métodos avaliados.

O gráfico evidencia, de forma clara, que **happy** é a emoção mais facilmente reconhecida, enquanto **disgust** e **fear** são consistentemente as mais difíceis. Esse padrão sugere que a extração de sinais visuais associados a nojo e medo é particularmente complexa para modelos baseados em linguagem multimodal, o que já é reportado na literatura de reconhecimento facial de emoções.

A tendência de sub-representação dessas classes aparece inclusive quando o modelo possui maior capacidade, como no caso do *qwen_instruct*, indicando que a dificuldade pode estar mais relacionada à natureza intrínseca das expressões e à distribuição do dataset do que ao modelo em si.

Comparação Entre Métodos

Para avaliar isoladamente o impacto do método de prompting, analisamos o F1 médio de cada classe agrupado por método. O gráfico seguinte contrasta diretamente como cada método se comporta em cada emoção.

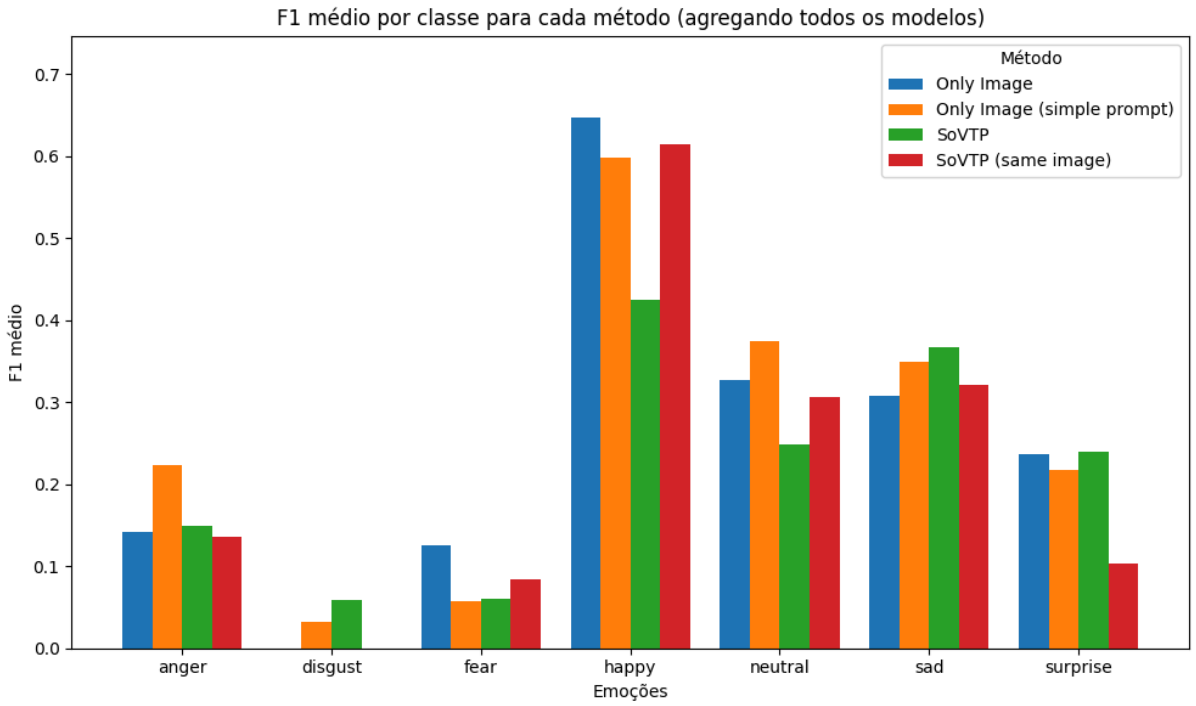


Figura 2 - F1-score médio por classe de emoção para cada método de prompting (Only Image, Only Image – simple prompt, SoVTP e SoVTP – same image), agregando os resultados de todos os modelos avaliados.

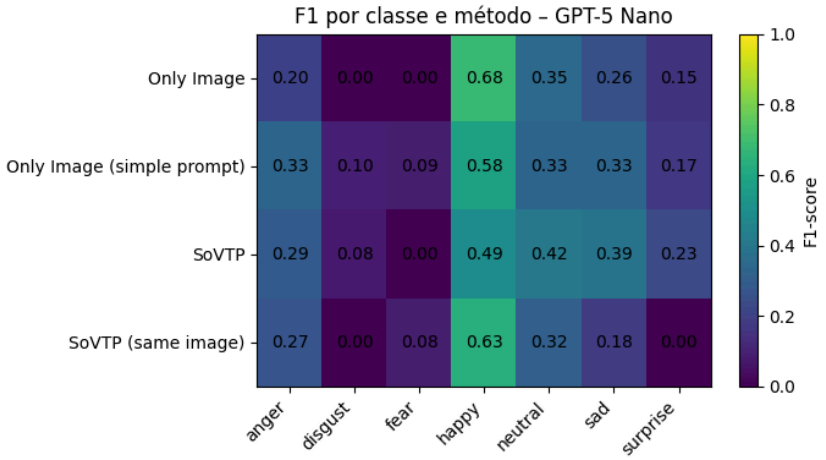
A partir dessa visualização, observa-se que:

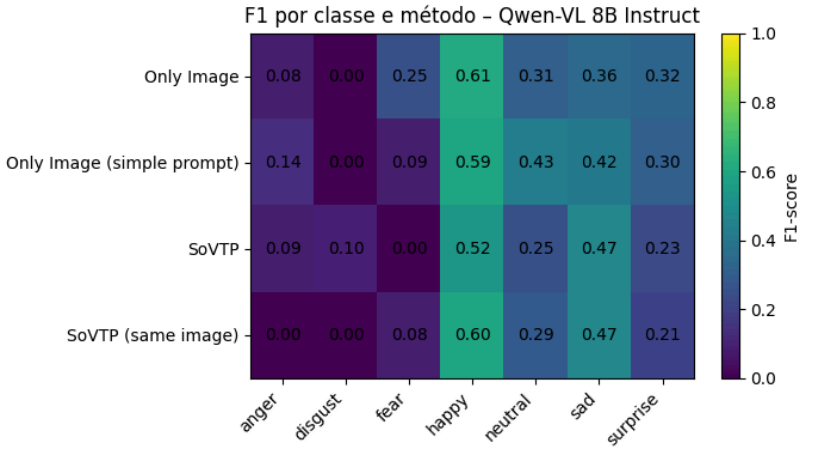
- **only_image_simple** apresenta os melhores valores de forma sistemática em grande parte das classes, configurando-se como o método mais forte e estável.
- **only_image** se destaca em **happy** e apresenta o melhor desempenho em **fear**, mesmo que o F1 ainda seja baixo.
- **sovtp**, apesar de teoricamente mais estruturado, não oferece vantagem clara na maioria das classes e parece trazer pequenos ganhos apenas em emoções negativas.
- **sovtp_sameimg** apresenta pior desempenho geral, reforçando a importância da construção progressiva de pistas visuais na versão original do método proposto por Wang et al. (2025).

Visualização por Heatmaps

Para destacar padrões mais finos de desempenho entre métodos e classes dentro de cada modelo, heatmaps foram utilizados. Esses mapas permitem identificar, de maneira imediata,

regiões de forte ou fraca performance e são úteis para capturar relações estruturais não triviais.





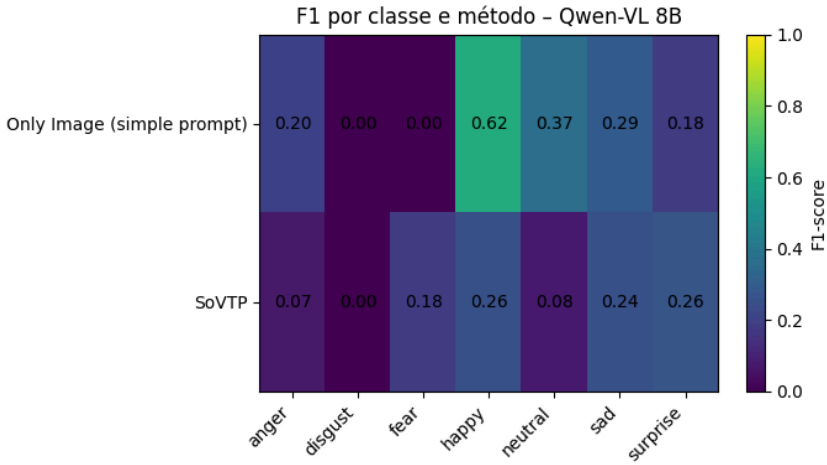


Figura 3 – Mapas de calor do F1-score por classe de emoção (anger, disgust, fear, happy, neutral, sad e surprise) para cada modelo avaliado (GPT-5 Nano, Qwen-VL 8B Instruct e Qwen-VL 8B), considerando os diferentes métodos de prompting em cada caso.

Esses heatmaps reforçam visualmente os padrões discutidos anteriormente: *disgust* e *fear* permanecem como emoções com desempenho quase nulo; *happy* aparece como o bloco mais claro; os métodos simples são superiores aos complexos na maioria dos casos; e *qwen_instruct* apresenta a maior consistência entre métodos.

Discussão Geral

Os resultados obtidos permitem concluir que MLLMs são capazes de detectar emoções básicas com desempenho acima do acaso, mas encontram limitações significativas em classes menos salientes ou menos frequentes. A superioridade do método **only_image_simple** sugere que, ao menos neste dataset, prompts diretos e minimalistas favorecem respostas mais estáveis.

A reprodução do método SoVTP, mesmo pautada pela proposta de Wang et al. (2025), não resultou em melhorias substanciais. Os ganhos observados em algumas emoções negativas não compensam a perda de precisão global, especialmente quando a versão adaptada sem imagens modificadas (**sovtp_sameimg**) é utilizada.

Além disso, a dificuldade extrema em *disgust* e *fear* sugere que a tarefa, tal como definida, depende fortemente de sinais visuais específicos que não foram plenamente capturados pelos modelos. Isso pode estar associado a limitações do dataset, à variabilidade das expressões, ou à própria abordagem zero-shot/ prompting adotada, indicando oportunidades claras para trabalhos futuros envolvendo técnicas supervisionadas, finetuning ou calibragem semântica de respostas.