cultureAI: Democratizando o Acesso a Museus com Large Language Models e Visualizações Imersivas Baseadas em Gaussian Splatting

Gabriel Valentim[1,2], **Giovanna M.**[1,3], **André C.**[1,2], **Luciano Soares**[2], **Fabio Ayres**[2]

[1]Nero.AI[‡]Insper - Instituto de Ensino e Pesquisa[2]UNESP - Universidade Estadual Paulista[3]

Resumo. *O Brasil enfrenta um cenário de exclusão cultural: 49% da população com nível fundamental nunca visitou um museu, 31% vive em cidades sem museus e 70% dos municípios carecem de equipamentos culturais. Apesar disso, 90% dos museus oferecem atividades educativas. O cultureAI surge como uma solução inclusiva para esse desafio, atuando como um curador virtual que leva o patrimônio cultural brasileiro (e latino americano) a comunidades historicamente marginalizadas. Combinando* Retrieval-augmented generation *(RAG) com visualizações 3D imersivas via* Gaussian Splatting*, o sistema permite explorar obras de arte com baixa latência, mesmo em smartphones simples.*

1. Introdução

O Brasil abriga mais de 3.800 museus, mas 49% dos adultos com com nível fundamental nunca visitaram um [ICOM Brasil 2021, IBGE 2023, IBRAM 2022]. A distância geográfica, os orçamentos limitados e a internet precária dificultam o acesso equitativo. Propomos o *cultureAI*, um chatbot com *Retrieval-augmented generation* (RAG) acoplado com a renderização dos modelos de *Gaussian Splatting* para democratizar o conteúdo museológico. Nossas contribuições incluem:

- uma pipeline escalável de RAG ajustada sobre 4.000 páginas de conteúdo de acervos de museus da américa latina (principalmente o MASP);
- um renderizador eficiente de *Gaussian Splatting* otimizado para dispositivos de baixo poder de processamento, uma vez que os modelos já estão treinados e públicos;
- utilização modular de *Large Language Models*, dando a possibilidade de executar o modelo diretamente no dispositivo.

2. Trabalhos Relacionados

Diante do contexto, alguns trabalhos foram cruciais para a criação de uma *pipeline* sustentável e enxuta. Em primeiro lugar, os avanços no campo das *NeRFs* trouxeram a possibilidade de renderizar cenas 3D complexas [Mildenhall et al. 2020], desde que fosse treinada sobre um dispositivo robusto. Porém, em contextos mais recentes, as técnicas de *Gaussian Splatting* [Kerbl et al. 2023] possibilitaram que a qualidade das mesmas cenas fossem maiores, bem como o hardware fosse menor. Isso trouxe escala para projetos que levasse em conta um número alto de instâncias a serem treinadas, armazenadas e renderizadas em dispositivos de baixo poder de processamento.

[‡]Nero.AI, São Paulo, Brasil. Este trabalho foi parcialmente desenvolvido durante o Llama Impact Pan-LATAM Hackathon 2024.

Além disso, vale ressaltar os avanços obtidos no campo dos modelos de linguagem. O recente lançamento do novo modelo *opensource Gemma3n* [DeepMind 2025] mostrou que mesmo em versões de poucos parâmetros (2B e 3B) é possível obter performance de modelos maiores. Isso pode ser evidenciado pelo fato dele ter alcançado 1300 no *benckmark LMArena* [Chiang et al. 2024], sendo o primeiro modelo a atingir esse valor com menos de 10B de parâmetros.

3. Execução em Dispositivo Imersivo

Para validar a portabilidade da solução, realizamos a execução do *cultureAI* em um *Meta Quest*, dispositivo de realidade virtual que também suporta experiências em realidade aumentada. A aplicação foi adaptada para rodar no navegador interno do dispositivo, permitindo que o usuário visualizasse algumas obras reconstruídas por meio do *Gaussian Splatting* em um ambiente totalmente imersivo.

Esse experimento evidencia a relação direta do trabalho com tecnologias de Realidade Aumentada (RA) e Realidade Virtual (RV). Ao integrar a renderização leve e de alta fidelidade provida pelo *Gaussian Splatting* a um headset acessível, torna-se possível enriquecer a experiência educacional: estudantes podem interagir com peças culturais em escala real, visualizar detalhes sob diferentes ângulos e trazer o acervo para dentro da sala de aula de forma imersiva. Dessa forma, a menção à RA não é apenas conceitual, mas fundamentada em um protótipo efetivamente executado em hardware de mercado.

4. Visão Geral do Sistema

A Figura 1 apresenta a arquitetura. Um backend em FastAPI orquestra ingestão de queries do usuário, aplicando o RAG clássico. Além disso, existe uma tarefa de classificar a intenção do usuário, para entender se ele fala sobre uma obra com a qual temos o modelo *splat* treinado. Sabendo que temos o modelo treinado, esse arquivo é mandado para um contêiner que renderiza a cena mostrando na tela do usuário tanto a resposta precisa do LLM, quanto o objeto 3D.

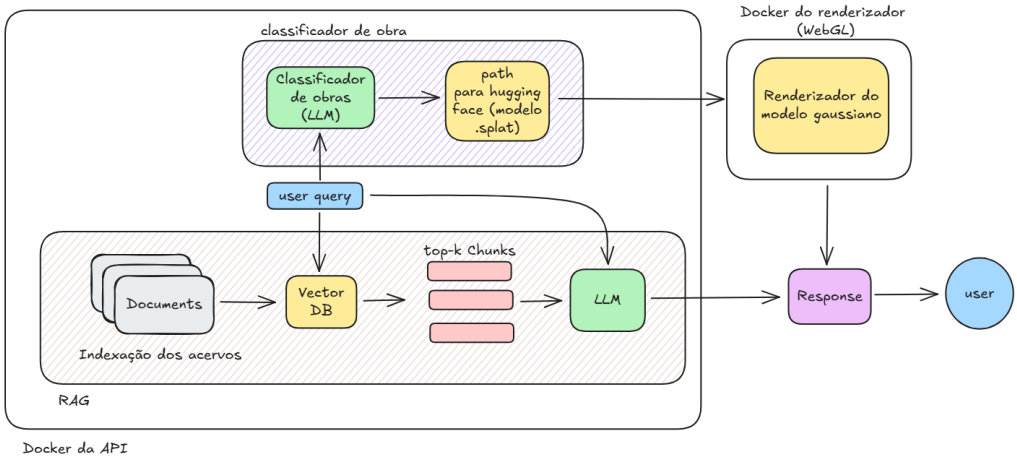

Figura 1. Pipeline do cultureAI combinando RAG e Gaussian Splatting.

4.1. Geração Aumentada por Recuperação

Cada consulta q é embutida em R^{1536} e os top-k trechos são concatenados ao prompt do LLM. Isso confere uma recuperação que ajuda a resposta do modelo a ser mais precisa do ponto de vista do conteúdo dos acervos.

5. Resultados Preliminares

O sistema foi implementado e testado em ambientes de competição (Top 6, *Meta - Llama Impact Hackathon Latam*), porém também possui lastro para teste em uma escala maior. Isso se deve, sobretudo, à otimização conferida pela técnica de *Gaussian Splatting*, bem como os avanços da performance de modelos *opensource*, minimizando o custo e conferindo escala.

6. Comparativo com Outras Representações

Em relação às malhas 3D tradicionais e nuvens de pontos, o *Gaussian Splatting* apresenta vantagens claras. Enquanto as malhas exigem etapas manuais de reconstrução, retopologia e texturização para alcançar realismo, o *splatting* deriva diretamente de fotografias com *Structure-from-Motion*, resultando em modelos leves e fiéis. Além disso, a rasterização diferenciável das gaussianas permite lidar melhor com transparências e efeitos dependentes do ponto de vista, algo mais custoso de reproduzir em malhas explícitas.

Comparado às abordagens baseadas em *NeRFs*, o *Gaussian Splatting* mantém a alta qualidade visual, mas com ganhos significativos em velocidade e portabilidade. Enquanto o *NeRF* clássico demanda longo tempo de treino e GPUs robustas, o *splatting* consegue renderização em tempo real mesmo em dispositivos modestos. Isso torna a técnica particularmente adequada para contextos educacionais e de inclusão cultural, onde acessibilidade e baixo custo de hardware são essenciais.

Vale ressaltar que estamos levando em consideração não só a performance da renderização, mas também a praticidade de treinamento que pode ser alcançada em dispositivos modestos. Isso promove um impacto cada vez mais positivo no contexto da democratização e do acesso à educação de qualidade por meio da imersão. Além de democratizar o acesso ao patrimônio cultural brasileiro e latino-americano, a solução demonstra potencial de aplicação em diferentes contextos e, principalmente, para enriquecer o processo de ensino-aprendizagem de estudantes, oferecendo uma alternativa acessível e escalável para o uso em ambientes educacionais.

7. Conclusão e Trabalhos Futuros

Para o contexto atual, fica clara a viabilidade técnica da resolução (mesmo que parcial) do problema de acessibilidade cultural. A combinação das tecnologias de estado da arte em visão computacional com o estado da arte dos modelos de linguagem de grande escala se torna uma ferramenta poderosa para o tratamento desse problema do mundo atual.

Por fim, urge a necessidade do desenvolvimento não só do trabalho feito até aqui, como também do total desacoplamento do sistema com a internet, isto é, inserir a solução num contexto *offline*. Devido aos resultados da manipulação de diferentes tamanhos de um mesmo vetor em [Kusupati et al. 2024], foi possível construir o que encontramos na arquitetura do modelo *Gemma3n* [DeepMind 2025]. A utilização das *MatFormer* [Devvrit et al. 2024] dá a capacidade de em um modelo maior, guardar modelos menores.

Referências

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot arena: An open platform for evaluating llms by human preference. In *arXiv preprint arXiv:2403.04132*.

DeepMind (2025). Gemma 3n. Link: https://ai.google.dev/gemma/docs/gemma-3n.

Devvrit, Kudugunta, S., Kusupati, A., Dettmers, T., Chen, K., Dhillon, I., Tsvetkov, Y., Hajishirzi, H., Kakade, S., Farhadi, A., and Jain, P. (2024). Matformer: Nested transformer for elastic inference. In *arXiv preprint arXiv:2310.07707*.

IBGE (2023). Sistema de Informações e Indicadores Culturais: 2011/2022. Instituto Brasileiro de Geografia e Estatística (IBGE), Coordenação de População e Indicadores Sociais, Períodos de referência de 01/01 a 31/12/2022. Link: https://www.ibge.gov.br/estatisticas/multidominio/cultura-recreacao-e-esporte/9388-indicadores-culturais.html.

IBRAM (2022). Pesquisa de Educação Museal no Brasil: Boletim Preliminar nº 1. Instituto Brasileiro de Museus (IBRAM) e Observatório da Economia Criativa da Bahia (OBEC), em colaboração com a Universidade Federal do Recôncavo da Bahia (UFRB) e Universidade Federal da Bahia (UFBA) 21 p., Link: ibram.

ICOM Brasil (2021). Pesquisa sobre museus e públicos no brasil - parte ii. In *Gov*, pages 8–9. Link: https://www.icom.org.br/wp-content/uploads/2021/01/ICOM_ES_Part-II.pdf.

Kerbl, B., Kopanas, G., Leimkuehler, T., and Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4).

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., and Farhadi, A. (2024). Matryoshka representation learning. In *arXiv preprint arXiv:2205.13147*.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*.