UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RONE CESAR BRANDÃO FILHO

De Imagem a Artefato: Geração de Objetos 3D Guiada por Texto para Experiências Interativas em Museus

Goiânia
2026

UFG

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 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 das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) 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 material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Rone Cesar Brandão Filho

3. Título do trabalho

De Imagem a Artefato: Geração de Objetos 3D Guiada por Texto para Experiências Interativas em Museus

4. 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) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
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;
- Publicação da dissertação/tese em livro.

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

Documento assinado eletronicamente por **Rone Cesar Brandao Filho**, **Discente**, em 23/06/2026, às 14:52, 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 **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 25/06/2026, às 11:16, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6247496** e o código CRC **9E647E25**.

Referência: Processo nº 23070.018420/2026-61 SEI nº 6247496

RONE CESAR BRANDÃO FILHO

De Imagem a Artefato: Geração de Objetos 3D Guiada por Texto para Experiências Interativas em Museus

Dissertação apresentada ao Programa de Pós–Graduação da Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação.
Linha de pesquisa: Sistemas Inteligentes e Aplicações.

Orientador: Prof. Arlindo Rodrigues Galvão Filho

Co-Orientador: Prof. Rafael Teixeira Sousa

Goiânia
2026

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

Brandão Filho, Rone Cesar

De Imagem a Artefato: Geração de Objetos 3D Guiada por Texto para Experiências Interativas em Museus [manuscrito] / Rone Cesar Brandão Filho. - 2026.

46 f.: 2026

Orientador: Prof. Dr. Arlindo Rodrigues Galvão Filho ; co-orientador: Dr. Rafael Texeira Sousa

Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2026.

Inclui: lista de figuras, lista de tabelas.

1. IA Generativa. 2. Reconstrução 3D. 3. Realidade Virtual. 4. Tours Imersivos em Museus. 5. Aprendizado Profundo.
I. Rodrigues Galvão Filho, Arlindo , orient. II. Texeira Sousa, Rafael, co-orient. III. Título.

CDU 004

UFG

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **9** da sessão de Defesa de Dissertação de **Rone Cesar Brandão Filho**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos sete dias do mês de maio de dois mil e vinte e seis, a partir das catorze horas, na sala 257 do INF, realizou-se a sessão pública de Defesa de Dissertação intitulada **"De Imagem a Artefato: Geração de Objetos 3D Guiada por Texto para Experiências Interativas em Museus"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Arlindo Rodrigues Galvão Filho (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Rafael Teixeira Andrade Sousa (UFMT), coorientador; Professor Doutor Rodrigo Zempulski Fanucchi (CEIA/AKCIT/UFG), membro titular externo; Professora Doutora Telma Woerle de Lima Soares, membra titular externa. A participação dos professores Rafael Teixeira Sousa e Rodrigo Zempulski Fanucchi ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Arlindo Rodrigues Galvão Filho, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos sete dias do mês de maio de dois mil e vinte e seis.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 07/05/2026, às 15:30, 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 **Telma Woerle De Lima Soares**, **Professora do Magistério Superior**, em 07/05/2026, às 17:23, 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 **Rone Cesar Brandao Filho**, **Discente**, em 07/05/2026, às 19:30, 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 **RODRIGO ZEMPULSKI FANUCCHI**, **Usuário Externo**, em 08/05/2026, às 10:02, 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 **Rafael Teixeira Sousa**, **Usuário Externo**, em 11/05/2026, às 10:34, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Referência: Processo nº 23070.018420/2026-61 SEI nº 6159906

Agradecimentos

Agradeço ao Prof. Arlindo Galvão e ao Dr. Rafael Teixeira pela orientação e apoio ao longo deste trabalho.

À minha esposa Giovanna, pelo companheirismo, pela paciência e pelo incentivo constante em todos os momentos desta jornada. Aos meus pais, Rone e Leila, ao meu irmão Ighor e ao meu primo Victor, pelo apoio incondicional.

Aos colegas de pesquisa do AKCIT, pela troca de experiências, pelo ambiente colaborativo e pelas contribuições ao longo deste trabalho.

À Universidade Federal de Goiás e ao Advanced Knowledge Center in Immersive Technologies (AKCIT), pelo suporte institucional e pela infraestrutura disponibilizada para o desenvolvimento desta pesquisa.

"O poder da educação em geral é pouco eficaz, exceto nas felizes ocasiões em que ele é quase supérfluo."

Edward Gibbon,
Declínio e queda do império romano.

Resumo

Brandão, Rone. **De Imagem a Artefato: Geração de Objetos 3D Guiada por Texto para Experiências Interativas em Museus**. Goiânia, 2026. 46p. Dissertação de Mestrado. Programa de Pós Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

Atualmente, muitos museus integram realidade virtual em seus roteiros, oferecendo experiências tanto virtuais quanto de realidade mista. Esse processo ainda está em estágios iniciais e costuma exigir profissionais qualificados e técnicas computacionais de alto custo, como fotogrametria, para gerar bons resultados. Nesse contexto, este trabalho apresenta uma abordagem simplificada que parte de um único panorama 360° e de um prompt em linguagem natural, produzindo modelos 3D texturizados adequados para aplicações de VR. O pipeline proposto combina segmentação guiada por prompt e modelos recentes de reconstrução 3D a partir de uma única imagem, eliminando fotogrametria e limpeza manual de malhas. Os artefatos reconstruídos foram integrados a um museu virtual interativo desenvolvido em Unity, enriquecido por um assistente de Geração Aumentada por Recuperação (RAG) capaz de responder perguntas dos visitantes em inglês com base em fontes curatoriais verificáveis. Uma avaliação preliminar com usuários demonstrou que os participantes conseguiram interagir com os artefatos e perceberam as informações fornecidas como suficientes para a compreensão dos objetos. Ao reduzir o esforço humano e a sobrecarga computacional, o pipeline proposto viabiliza experiências de museu mais interativas, imersivas e facilmente atualizáveis para instituições de qualquer porte.

Palavras–chave

<IA Generativa, Reconstrução 3D, Realidade Virtual, Realidade Mista, Tours Imersivos em Museus, Aprendizado Profundo, Digitalização de Patrimônio Cultural>

Abstract

Brandão, Rone. **From Image to Artifact: Text-Guided 3D Object Generation for Interactive Museum Experiences**. Goiânia, 2026. 46p. MSc. Dissertation. Programa de Pós Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

Currently, many museums are integrating virtual reality into their tours, offering both virtual and mixed reality experiences. This process is still in its early stages and often requires skilled professionals and high-cost computational techniques such as photogrammetry to generate good results. In this context, this work presents a streamlined approach that starts from a single 360° panorama and a natural language prompt, producing textured 3D models suitable for VR applications. The proposed pipeline combines prompt-guided segmentation with recent single-image 3D reconstruction models, eliminating photogrammetry and manual mesh cleanup. The reconstructed artifacts were integrated into an interactive virtual museum developed in Unity, enriched by a Retrieval-Augmented Generation (RAG) assistant capable of answering visitor questions in English based on verifiable curatorial sources. A preliminary user evaluation demonstrated that participants were able to interact with the artifacts and perceived the information provided as sufficient for understanding the objects. By reducing human effort and computational overhead, the proposed pipeline enables more interactive, immersive, and easily updatable museum experiences for institutions of any size.

Keywords

<Generative AI, 3D Reconstruction, Virtual Reality, Mixed Reality, Immersive Museum Tours, Deep Learning, Cultural Heritage Digitization>

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

A forma como um museu expõe seus artefatos desempenha um papel fundamental na formação da experiência do visitante [25]. Nos últimos anos, esforços significativos têm sido empreendidos na digitalização 3D do patrimônio cultural, aprimorando a maneira como as instituições apresentam suas coleções e permitindo que o público desfrute de experiências mais interativas e imersivas. Tecnologias imersivas possibilitam que os visitantes interajam com peças de valor inestimável, oferecendo não apenas informações textuais, mas também uma compreensão mais profunda e uma visualização aprimorada dos objetos. Além disso, alguns artefatos podem ser extremamente frágeis ou valiosos para exposição, e certos objetos históricos podem já não existir ou apresentar danos que inviabilizem sua apresentação [2]. Como os museus normalmente exibem apenas uma pequena parte de suas coleções [11], a realidade virtual (RV) possibilita o acesso a um leque mais amplo de artefatos, transformando o ambiente museológico de um espaço estático e limitado em um espaço dinâmico e envolvente [30, 2]. Nesse cenário, a RV oferece aos pesquisadores uma forma aprimorada de comunicar resultados científicos, tanto para a comunidade acadêmica quanto para o público em geral [25, 16].

Apesar do reconhecido potencial da digitalização 3D para a preservação e disseminação do patrimônio cultural, o método predominante na prática, a fotogrametria, impõe barreiras consideráveis à sua adoção em larga escala. A reconstrução fotogramétrica de um único artefato exige tipicamente dezenas a centenas de fotografias capturadas em ângulos controlados, equipamentos especializados (tripés, iluminação difusa, fundos neutros), operadores treinados e um extenso fluxo de pós-processamento que inclui alinhamento de nuvens de pontos, limpeza de malha e reparo de texturas [24]. Esse processo pode consumir horas ou mesmo dias por objeto, tornando a digitalização de acervos com milhares de peças economicamente inviável para a maioria das instituições [27]. Consequentemente, grande parte do patrimônio cultural mundial permanece inacessível ao público digital, restrita a reservas técnicas sem qualquer representação tridimensional [14]. Identifica-se, portanto, a necessidade de abordagens alternativas que reduzam drasticamente o custo, o tempo e a expertise técnica requeridos para a geração de ativos 3D a partir de acervos museológicos.

Avanços recentes em modelos generativos a partir de uma única imagem oferecem uma perspectiva promissora para essa questão. Tecnologias de síntese de novas vistas, como campos de radiância neural (NeRFs) [15] e 3D Gaussian Splatting (3DGS) [9], têm permitido a geração de representações 3D completas a partir de poucas ou mesmo de uma única imagem [39, 20, 1]. Combinados com modelos de difusão [12], esses métodos evoluíram significativamente, conseguindo gerar conteúdo 3D realista e texturizado em um curto espaço de tempo [29]. Entretanto, apesar dos progressos em qualidade de renderização, métodos baseados em NeRFs ainda apresentam treinamento e renderização dispendiosos em termos de tempo e de hardware [36], enquanto técnicas fundamentadas em 3DGS têm sua qualidade prejudicada ao lidar com entradas desafiadoras, como cenários com vistas esparsas, efeitos de sombreamento complexos e cenas de grande escala [21].

Embora os modelos generativos recentes tenham demonstrado capacidade de produzir objetos 3D de alta qualidade a partir de uma única imagem, sua aplicação direta a ambientes museológicos reais introduz desafios adicionais que permanecem não resolvidos. Em primeiro lugar, um único panorama 360° de uma sala de museu pode conter dezenas de artefatos heterogêneos, muitos dos quais são visualmente semelhantes ao fundo circundante, exigindo que cada objeto seja isolado individualmente antes da reconstrução. Em segundo lugar, as condições de captura em museus envolvem iluminação não controlada, reflexos em vitrines de vidro e oclusões parciais entre objetos adjacentes. Em terceiro lugar, modelos de segmentação convencionais com classes predefinidas [5, 3] exigem rótulos limitados e extensas anotações por item, mostrando-se ineficazes para o vocabulário aberto e diversificado de artefatos museológicos [26]. Por fim, não se identifica na literatura um pipeline que conecte a cadeia completa, desde a imagem panorâmica até a segmentação guiada por texto, a reconstrução 3D e a integração em uma cena de RV interativa com informação contextual.

Diante dessa lacuna, o presente trabalho propõe um pipeline de duas etapas que, a partir de um único panorama 360° e um prompt em linguagem natural, segmenta artefatos individuais e reconstrói objetos 3D texturizados adequados a aplicações de realidade virtual. A abordagem elimina a necessidade de captura multi-vista, iluminação controlada, operadores especializados ou modelagem manual, requerendo apenas uma imagem panorâmica já disponível em passeios virtuais de museus. Para a segmentação, adota-se o Grounded-SAM-2 [7], que combina o Grounding DINO [13] para detecção guiada por texto com o SAM 2 [22] para refinamento de máscaras a nível de pixel. Para a reconstrução, comparam-se três modelos estado da arte, a saber, Trellis [37], Hunyuan3D 2.1 [6] e Unique3D [35], e o modelo selecionado é empregado na construção de uma cena de museu virtual interativa em Unity. Complementarmente, integra-se uma camada de Geração Aumentada por Recuperação (RAG) que permite aos visitantes consultar

informações contextuais sobre os artefatos por meio de linguagem natural, com respostas geradas em inglês pelo modelo GPT-4o mini [17].

As principais contribuições deste trabalho são:

1. Uma abordagem de segmentação guiada por prompt baseada no Grounded-SAM-2, capaz de isolar artefatos individuais em panoramas complexos de museus sem classes predefinidas ou anotação manual.
2. Uma comparação qualitativa de três modelos de reconstrução 3D a partir de imagem única (Trellis, Hunyuan3D 2.1 e Unique3D), aplicados a artefatos museológicos em condições reais de iluminação e enquadramento.
3. Um pipeline completo e de baixo custo que conecta a cadeia panorama $\rightarrow$ segmentação $\rightarrow$ reconstrução 3D $\rightarrow$ cena de museu virtual interativa, a partir de uma única imagem de entrada por objeto.
4. A integração de um assistente conversacional baseado em RAG, que fornece informações contextuais sobre os artefatos em inglês, fundamentadas em fontes curatoriais verificáveis.
5. Uma avaliação preliminar com usuários, mensurando recordação de artefatos, suficiência informacional, grau de imersão e usabilidade da interação.

O restante desta dissertação está organizado da seguinte forma. O Capítulo 2 apresenta os trabalhos relacionados, abrangendo digitalização de patrimônio cultural, geração 3D a partir de imagens, segmentação guiada por texto e assistentes baseados em modelos de linguagem. O Capítulo 3 descreve os materiais e métodos empregados, detalhando o conjunto de dados, o pipeline de segmentação e reconstrução, a construção da cena em Unity e a implementação do assistente RAG. O Capítulo 4 apresenta e discute os resultados obtidos, incluindo a comparação entre os modelos de reconstrução e a avaliação com usuários. Por fim, o Capítulo 5 sintetiza as conclusões, reconhece as limitações do trabalho e aponta direções para pesquisas futuras.

CAPÍTULO 2

Trabalhos Relacionados

O presente capítulo organiza a literatura relevante em três eixos temáticos: (i) digitalização e virtualização do patrimônio cultural, com ênfase nas limitações de escala e custo das abordagens tradicionais; (ii) geração 3D a partir de imagens, com análise comparativa das arquiteturas mais recentes; e (iii) assistentes inteligentes e geração aumentada por recuperação (RAG) aplicados a museus. O capítulo encerra posicionando o presente trabalho em relação às lacunas identificadas.

2.1 Digitalização e Virtualização do Patrimônio Cultural

A digitalização de acervos museológicos tem sido impulsionada pela necessidade de ampliar o acesso público a coleções que, em sua maioria, permanecem fora das galerias físicas [11]. Nesse contexto, a fotogrametria consolidou-se como técnica dominante por produzir réplicas digitais geometricamente precisas a partir de sequências fotográficas densas [2]. Contudo, como demonstram Carvajal et al. [2], o fluxo de trabalho fotogramétrico depende fortemente de operadores especializados, controle rigoroso de iluminação e extenso pós-processamento manual, o que limita sua escalabilidade para instituições com acervos extensos ou recursos técnicos reduzidos.

Roussou [25] apresenta uma das primeiras aplicações sistemáticas de VR para difusão cultural, conduzida pela *Foundation of the Hellenic World* (FHW). O trabalho demonstra que ambientes imersivos aumentam a memorização de fatos históricos em comparação a exposições tradicionais. No entanto, a produção exigiu anos de modelagem manual e equipes multidisciplinares, evidenciando que a qualidade da experiência imersiva estava diretamente atrelada ao volume de trabalho humano empregado. Essa dependência de esforço manual intensivo permanece uma limitação central em grande parte dos sistemas subsequentes.

Pantile et al. [19] avançam nessa direção ao introduzir o conceito de *visit-actor*, um modelo de visitante ativo que interage com réplicas virtuais sobrepostas ao espaço físico por meio de AR e VR. A avaliação qualitativa reporta maior engajamento dos usuários; contudo, os autores reconhecem que a criação de modelos 3D fotorrealistas

dentro dos prazos exigidos por exposições temporárias constitui um gargalo crítico. Essa tensão entre qualidade visual e agilidade de produção é recorrente na literatura e motiva diretamente a busca por pipelines generativos mais rápidos.

Orr et al. [18], em revisão sistemática de 48 estudos de caso sobre aprendizagem informal em museus com XR, identificam três fatores críticos para experiências significativas: fidelidade visual, interatividade adaptativa e integração curricular. Relevante para o presente trabalho é a constatação de que falhas na reconstrução geométrica prejudicam a presença cognitiva do usuário, sugerindo que a qualidade dos modelos 3D gerados impacta diretamente os resultados de aprendizagem e não apenas a estética da experiência.

Magnani et al. [14] investigam a fotogrametria participativa aplicada a coleções indígenas Sámi, em que membros da comunidade operam câmeras e softwares para reconstruir objetos cerimoniais. Embora a abordagem fortaleça a agência comunitária e produza modelos culturalmente significativos, ela ainda exige múltiplas capturas fotográficas sob condições controladas de iluminação, algo inviável em coleções dispersas ou em artefatos expostos em vitrines com reflexos especulares. Essa limitação é particularmente relevante para o cenário explorado neste trabalho, onde os objetos são capturados em condições reais de museu, frequentemente subótimas.

Em conjunto, esses trabalhos evidenciam que as abordagens tradicionais de digitalização, embora produzam resultados de alta fidelidade, compartilham limitações estruturais: dependência de captura multi-view controlada, alto custo de mão de obra especializada e baixa escalabilidade. Essas lacunas motivam a investigação de alternativas baseadas em modelos generativos capazes de operar a partir de entradas mínimas, como uma única imagem extraída de uma panorâmica de 360°.

2.2 Geração 3D a partir de Imagens

A geração automática de objetos 3D a partir de imagens 2D representa uma das fronteiras mais ativas da visão computacional e da computação gráfica. Os avanços nessa área podem ser organizados em três gerações de abordagens: métodos baseados em campos de radiância neural (NeRFs), técnicas de *Gaussian Splatting* e, mais recentemente, modelos generativos de difusão orientados à reconstrução de malhas texturizadas.

2.2.1 Campos de Radiância Neural e Gaussian Splatting

Mildenhall et al. [15] introduziram os NeRFs como uma representação implícita de cenas 3D, em que uma rede neural *multilayer perceptron* (MLP) mapeia coordenadas espaciais e direções de visão para valores de densidade volumétrica e radiância. Essa abordagem produziu avanços expressivos na síntese de novas vistas, mas o treinamento

e a renderização permanecem dispendiosos em termos de tempo e hardware [36], o que limita sua aplicabilidade em cenários que demandam processamento rápido, como a digitalização em larga escala de acervos museológicos.

O 3D Gaussian Splatting (3DGS), proposto por Kerbl et al. [9], contorna parte dessas limitações ao representar cenas como distribuições gaussianas projetáveis, permitindo renderização em tempo real com taxas entre 100 e 300 quadros por segundo em GPUs de alto desempenho [36]. Contudo, como apontam Radl et al. [21], a qualidade do 3DGS degrada-se sensivelmente em cenários com vistas esparsas, efeitos de sombreamento complexos e cenas de grande escala, condições frequentes em ambientes de museu capturados por panorâmicas de 360°. Ademais, tanto NeRFs quanto 3DGS pressupõem múltiplas imagens do objeto como entrada, o que os torna inadequados para o cenário de imagem única explorado neste trabalho.

2.2.2 Modelos Generativos de Difusão para Reconstrução 3D

A limitação de entrada multi-view motivou o desenvolvimento de modelos generativos capazes de sintetizar representações 3D a partir de uma única imagem. Liu et al. [12] demonstraram que modelos de difusão latente pré-treinados podem ser adaptados para gerar vistas sintéticas de um objeto a partir de uma única fotografia, abrindo caminho para pipelines de reconstrução baseados em visão única. Shi et al. [28] avançaram nessa direção com o Zero123++, incorporando mosaico de vistas supervisionadas e condicionamento por atenção escalonada para produzir múltiplas vistas geometricamente consistentes sem a necessidade de estimadores externos de pose, alcançando maior coerência geométrica entre as imagens geradas em comparação ao trabalho anterior.

Sobre essa base, Wu et al. [35] propõem o Unique3D, um pipeline ponta a ponta que integra o Zero123++ para geração multivista e introduz o algoritmo *Instant and Consistent Mesh Reconstruction* (ISOMER). O diferencial está na *loss* de orientação de vértices baseada em *normal maps*, que estabiliza o otimizador durante o colapso e divisão de arestas, resultando em ganhos de 18% em LPIPS no dataset CO3D. Apesar da alta fidelidade superficial, o Unique3D tende a produzir geometria com ruído de alta frequência e buracos ocasionais, especialmente em regiões de baixa textura, como observado nos experimentos deste trabalho.

Xiang et al. [37] apresentam o Trellis, treinado em até 2 bilhões de parâmetros, que introduz a representação *Structured LATent* (SLat): uma grade 3D esparsa que associa vetores latentes locais a voxels ativos, codificados por um *Variational Autoencoder* (VAE) baseado em Transformer sobre características DINOv2. A geração opera em duas etapas com *Rectified Flow Transformer*: o primeiro modelo gera a estrutura esparsa da grade e o segundo infere os vetores latentes de cada célula não vazia. O espaço SLat pode ser

decodificado em diferentes representações finais, como 3D Gaussians, Radiance Fields ou malhas FlexiCubes, conferindo versatilidade de formato. Em comparação ao Unique3D, o Trellis produz geometria mais limpa e sem auto-interseções, porém com menor fidelidade a detalhes locais de alta frequência.

Yang et al. [6] apresentam o Hunyuan3D 2.1, sistema que dissocia forma e aparência em dois módulos coordenados. O módulo de geometria, baseado em *flow matching*, gera uma malha alinhada à imagem de entrada por meio de um ShapeVAE que comprime a geometria em tokens latentes com amostragem de importância em bordas e cantos. O módulo de texturização recebe mapas de posição e normais da malha, remove a iluminação e injeta características da imagem original para produzir materiais em alta resolução compatíveis com o padrão PBR (*Physically Based Rendering*), prontos para uso em motores gráficos como Unity e Unreal. Em comparação ao Trellis e ao Unique3D, o Hunyuan3D 2.1 apresenta o melhor equilíbrio entre fidelidade geométrica, qualidade de texturização e tempo de inferência para o cenário de entrada única, o que motivou sua adoção como modelo de reconstrução na etapa final do pipeline proposto neste trabalho, após comparação qualitativa com os demais modelos.

Em síntese, os modelos de difusão para reconstrução 3D avançaram consideravelmente na qualidade geométrica e na velocidade de inferência, tornando viável a digitalização a partir de imagens únicas. Contudo, todos os métodos revisados pressupõem que o objeto de interesse já foi isolado da cena, não abordando o problema de segmentação em panorâmicas densas típicas de galerias de museus. Essa lacuna é endereçada pelo pipeline proposto neste trabalho, que integra segmentação guiada por texto à reconstrução 3D de imagem única.

2.3 Assistentes Inteligentes e RAG em Museus

A crescente adoção de modelos de linguagem de grande escala (LLMs) em aplicações culturais e educacionais abriu novas possibilidades para a mediação do conhecimento em museus virtuais. Diferentemente das abordagens tradicionais baseadas em painéis textuais estáticos ou audioguias pré-gravados, sistemas baseados em LLMs permitem interações conversacionais dinâmicas, personalizadas e adaptadas ao perfil e às perguntas específicas de cada visitante [4].

Vasic et al. [33] propõem um guia museológico personalizado baseado em LLM, em que o sistema adapta o roteiro da visita às preferências declaradas pelo usuário, como período histórico de interesse ou nível de conhecimento prévio. O sistema foi avaliado em um museu de arte com resultados positivos em satisfação e engajamento. Contudo, os autores identificam um risco central: sem mecanismos de ancoragem ao acervo real, o modelo tende a produzir respostas plausíveis porém factualmente incorretas, fenômeno

conhecido como alucinação. Essa limitação motiva diretamente o uso de técnicas de geração aumentada por recuperação.

Wang et al. [34] apresentam o VirtuWander, sistema de guia virtual para tours panorâmicos que combina LLMs com interação multimodal. O sistema permite que usuários apontem para regiões de interesse em imagens de 360° e recebam explicações contextuais geradas pelo modelo. Embora o VirtuWander demonstre a viabilidade de integrar LLMs a ambientes panorâmicos, ele opera sobre imagens estáticas e não constrói representações 3D dos objetos, limitando a profundidade da experiência imersiva em comparação a ambientes de VR com geometria reconstruída.

A geração aumentada por recuperação (RAG) surge como resposta direta ao problema de alucinação, ancorando as respostas do LLM em fontes curadas e verificáveis [4]. Em domínios como patrimônio cultural, onde factualidade, proveniência e sensibilidade cultural são críticas, o RAG é particularmente relevante: ao invés de depender exclusivamente do conhecimento paramétrico do modelo, o sistema recupera passagens relevantes de uma base documental controlada, como fichas catalográficas, textos curatoriais e referências bibliográficas, e as fornece como contexto para a geração da resposta. Cossatin et al. [4] demonstram essa abordagem em um website de patrimônio cultural, reportando aumento na precisão factual e na confiança dos usuários nas respostas geradas.

Lau et al. [10] avaliam a usabilidade e o engajamento de assistentes baseados em LLM integrados a ambientes de VR em um estudo com o esporte tradicional escocês de curling. Os resultados indicam ganhos mensuráveis em usabilidade e engajamento quando chatbots conversacionais são empregados em lugar de interfaces de menu tradicionais, sugerindo que a modalidade conversacional é mais natural para usuários em imersão. Esse achado é relevante para o presente trabalho, que adota interação por apontamento (*raycast*) combinada a respostas geradas por RAG como mecanismo principal de mediação do conhecimento na cena VR.

Em contraste com os trabalhos revisados, que tratam a entrega de conteúdo e a criação de ativos 3D como problemas separados, o presente trabalho integra segmentação guiada por texto, reconstrução 3D de imagem única e RAG orientado por artefato em um pipeline unificado. A camada RAG é implementada com o modelo GPT-4o mini [17], operando sobre uma base de conhecimento extraída do site oficial do Museu Egípcio do Cairo, com respostas fornecidas em inglês. Quando o usuário aponta para um artefato na cena Unity, um *raycast* recupera a chave única κ do objeto; o sistema então consulta um índice vetorial denso, recupera as passagens mais relevantes e as fornece como contexto ao modelo, que gera uma resposta ancorada ao visitante.

2.4 Considerações Finais

A revisão da literatura evidencia três lacunas inter-relacionadas que o presente trabalho se propõe a abordar. Primeiro, as abordagens tradicionais de digitalização museológica, como fotogrametria e modelagem manual, produzem resultados de alta fidelidade, mas são inescaláveis para instituições com acervos extensos ou recursos técnicos limitados [2, 14]. Segundo, os modelos generativos de reconstrução 3D mais recentes, embora capazes de operar a partir de imagens únicas, pressupõem que o objeto de interesse já foi isolado da cena, não abordando o problema de segmentação em panorâmicas densas típicas de galerias [37, 35, 6]. Terceiro, os sistemas de mediação baseados em LLM e RAG para museus virtuais existentes operam sobre imagens estáticas ou ambientes sem geometria 3D reconstruída, limitando a profundidade da experiência imersiva [33, 34].

O pipeline proposto neste trabalho fecha essas lacunas ao integrar, em uma única cadeia automatizada: (i) segmentação aberta guiada por texto a partir de panorâmicas de 360°; (ii) reconstrução 3D de imagem única com comparação entre Trellis, Unique3D e Hunyuan3D 2.1; e (iii) um assistente RAG orientado por artefato, integrado a uma cena VR interativa construída em Unity, capaz de responder perguntas dos visitantes em inglês com base em fontes curatoriais verificáveis.

CAPÍTULO 3

Materiais e Métodos

Nesta seção, descreve-se o conjunto de dados empregado na avaliação da abordagem e apresenta-se uma visão detalhada do pipeline proposto em duas etapas. O pipeline consiste em: (1) uma etapa de detecção e segmentação de objetos baseada em prompts, na qual artefatos individuais são isolados de cenas 360° complexas por meio de prompts guiados por texto; e (2) uma etapa de reconstrução de objetos 3D, na qual cada artefato segmentado é convertido em um objeto 3D texturizado a partir de um único recorte mascarado.

3.1 *Dataset*

O objetivo deste conjunto de dados é avaliar um cenário de tour virtual real, com objetos capturados em condições desfavoráveis de ângulo e iluminação. Para a comparação qualitativa entre os modelos de geração 3D, selecionaram-se quadros de vídeos 360° de três museus: Egyptian Museum in Cairo, Metropolitan Museum of Art em Nova York e National Museum of Natural History nas Filipinas. A partir desses vídeos, foram escolhidos aleatoriamente objetos como esculturas, jarros, pinturas, fósseis e outros, com a visualização completa do objeto.

Para a composição do museu virtual interativo, utilizou-se exclusivamente o acervo do Museu Egípcio do Cairo (MEC), disponível em vídeos de tour 360° online, conforme ilustrado na Figura 3.1. A restrição a um único museu teve como objetivo controlar a variabilidade visual e garantir consistência nos artefatos selecionados, além de viabilizar a curadoria das informações catalográficas necessárias para a base de conhecimento do assistente RAG, extraídas do site oficial do MEC [31].

3.2 Segmentação Guiada por Prompt

A escolha do Grounded-SAM-2 como módulo de detecção e segmentação fundamenta-se em duas limitações dos detectores convencionais. Primeiro, métodos como

Figura 3.1: *Exemplo de três quadros selecionados aleatoriamente de tours online.*

YOLO [23] e Mask R-CNN [5] operam com vocabulário fechado, isto é, reconhecem apenas as categorias definidas no conjunto de treinamento. Artefatos museológicos apresentam diversidade tipológica elevada, abrangendo esculturas, relevos, vasos, mobiliário e joias, categorias que não correspondem diretamente às classes desses datasets. Adaptar um detector de vocabulário fechado exigiria coleta de dados anotados e retreinamento para cada nova categoria. O Grounded-SAM-2 contorna essa restrição ao operar com vocabulário aberto: o Grounding DINO aceita prompts em linguagem natural, permitindo descrever qualquer artefato sem anotação prévia nem retreinamento. Segundo, detectores como YOLO produzem apenas caixas delimitadoras, não máscaras a nível de pixel. Para gerar entradas limpas para o modelo de reconstrução 3D, isto é, recortes RGBA sem fundo, a segmentação com precisão de pixel proporcionada pelo SAM 2 é essencial. A integração de detecção e segmentação em pipeline único elimina etapas manuais intermediárias.

Dado um panorama equiretangular I e uma consulta de prompt do usuário q, o pipeline passa para a etapa de detecção, utilizando o Grounding DINO, que recebe (I, q) como entrada e devolve um conjunto de caixas delimitadoras $\{b_i\}$ alinhadas ao prompt textual. Cada caixa b_i que permanece é então refinada pelo SAM 2, o qual produz uma máscara m_i de alta resolução e precisão de pixel.

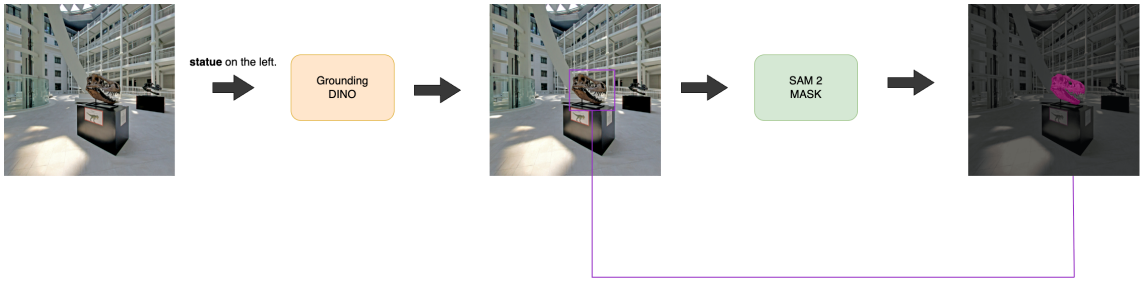

Figura 3.2: *Etapa de segmentação guiada por prompt. O panorama e o prompt são inicialmente enviados ao Grounding DINO, que devolve uma caixa delimitadora condicionada ao texto. Essa caixa é encaminhada ao SAM 2, cujo decodificador de máscara irá refinar e retornar uma máscara de segmentação a nível de pixel.*

Os prompts utilizados neste trabalho seguem linguagem natural livre, sem necessidade de sintaxe especial ou vocabulário controlado. O usuário pode referenciar o objeto diretamente pelo seu tipo (*e.g.*, `"estátua"`, `"jarro"`, `"fóssil"`), ou complementar a descrição com informações de posicionamento e direção espacial para desambiguar artefatos semelhantes presentes na mesma cena (*e.g.*, `"estátua à esquerda"`, `"escultura ao fundo"`, `"vaso no centro"`). Em todos os artefatos avaliados sob condições adequadas de iluminação e visibilidade, o pipeline foi capaz de identificar corretamente o objeto de interesse a partir do prompt fornecido, sem que nenhum item testado deixasse de ser detectado. Cenários adversos, como artefatos posicionados atrás de vitrines com reflexos especulares ou em ambientes de baixa luminosidade, introduzem degradações na qualidade da máscara e são discutidos como limitações na Seção 4.1.1.

Conforme ilustrado na Figura 3.2, o Grounding DINO incorpora primeiramente a consulta textual aos tokens do panorama, de forma que cada proposta de objeto seja filtrada semanticamente antes de receber pontuação. Neste exemplo, o detector gera três caixas candidatas, mas apenas aquela cuja confiança excede o limiar sobrevive à supressão de não máximos. Ao encaminhar essa única caixa ao SAM 2 evita-se a supersegmentação que frequentemente ocorre quando toda a imagem é mascarada de uma só vez.

O Transformer SAM 2, por ser leve, processa apenas os pixels internos ou imediatamente adjacentes à caixa delimitadora, reduzindo o tempo de inferência e ajudando a recuperar contornos finos, como as bordas irregulares dos dentes do fóssil. A máscara binária resultante é multiplicada pelo panorama para criar um recorte RGBA (512×512), que preserva a iluminação e a textura originais, descartando os pixels de fundo. Esse recorte torna-se a única entrada do gerador 3D, garantindo que a etapa de reconstrução se concentre no artefato e não no salão ao redor.

Adotaram-se os parâmetros padrão do demo oficial do Grounded-SAM-2 [7] sem ajuste adicional. Para o Grounding DINO, utilizou-se o modelo com *backbone* SwinT (GroundingDINO_SwinT_OGC), com limiar de caixa (*box_threshold*) de 0,35 e limiar de texto (*text_threshold*) de 0,25. Para o SAM 2, empregou-se o modelo SAM 2.1

Hiera Large, com a opção de saída multimáscara desativada (*multimask_output* = False), de modo que o decodificador retorna apenas a máscara de maior confiança. Diferentes formulações de prompt foram testadas informalmente ao longo do desenvolvimento, variando entre descrições genéricas (`"estátua"`) e especificações espaciais (`"estátua à esquerda"`), confirmando robustez qualitativa do detector a variações na formulação. Cabe ressaltar que a avaliação da segmentação foi observacional, sem um dataset anotado de referência.

3.3 Geração 3D

A avaliação concentra-se em três modelos generativos recentes, selecionados segundo quatro critérios: (1) representatividade arquitetural, de modo a cobrir paradigmas distintos de geração, a saber, difusão multivista (Unique3D), *Rectified Flow* com latentes estruturados (Trellis) e *Flow Matching* com *Diffusion Transformer* (Hunyuan3D 2.1); (2) suporte nativo a entrada de imagem única, requisito imposto pelo pipeline proposto; (3) disponibilidade de código-fonte e pesos abertos, permitindo reprodutibilidade; e (4) desempenho competitivo reportado na literatura recente em benchmarks de reconstrução 3D.

Durante a fase exploratória, outros modelos foram avaliados preliminarmente. O TripoSR [32] e o InstantMesh [38], ambos baseados em *Large Reconstruction Models*, foram testados com recortes de artefatos museológicos, porém apresentaram qualidade geométrica e de textura visivelmente inferior à dos três modelos selecionados, especialmente em regiões com detalhes finos e em superfícies curvas. Esses resultados motivaram a exclusão de ambos da comparação formal.

A comparação entre os três modelos adota uma abordagem qualitativa centrada na percepção visual, dado que o objetivo final é a composição de uma experiência interativa em realidade virtual. Nesse contexto, a adequação ao cenário VR depende de atributos como qualidade percebida de texturas, coerência de silhuetas e hermeticidade da malha, mais do que precisão geométrica absoluta. Além disso, não existem modelos 3D de referência (*ground truth*) dos artefatos museológicos avaliados, o que impossibilita o cálculo de métricas como *Chamfer Distance* ou *F-score*. Os critérios de avaliação visual e os resultados da comparação são detalhados no Capítulo 4.

3.3.1 Trellis

Trellis é uma família de modelos generativos 3D concebida para criar ativos de alta qualidade a partir de texto ou de imagens condicionantes, mantendo compatibilidade com diferentes formatos de saída (Gaussians, Radiance Fields ou malhas). O pilar concei-

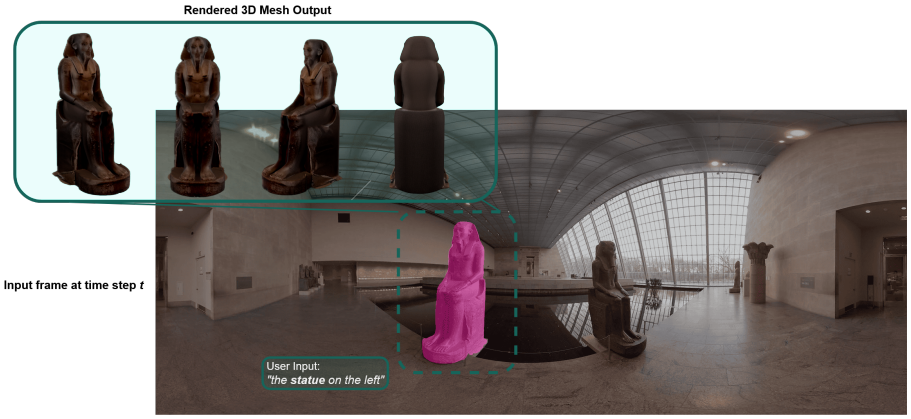

Figura 3.3: *Visualização de um objeto reconstruído de um panorama 360°, e exibido em diferentes perspectivas.*

tual da abordagem é o Structured LATent (SLAT), uma grade esparsa em três dimensões cujos voxels ativos, isto é, aqueles que interceptam a superfície do objeto, armazenam vetores latentes locais capazes de codificar simultaneamente geometria e aparência. Como a ocupação da grade cresce apenas com a área da superfície, a representação preserva resolução sem penalizar memória e permite reamostragem local fina durante a edição. Além disso, por manter a mesma semântica em todo o pipeline, o SLAT pode ser decodificado por cabeças diferentes, cada qual especializada em um tipo de representação final, sem exigir *refitting* ou afinamento adicional.

Para aprender esse espaço latente, cada objeto de treino é renderizado em 150 imagens RGB a 512x512 px distribuídas uniformemente em ângulos polares. Essas projeções são processadas por um backbone DINO-v2 congelado, cujos tokens visuais são projetados nos voxels correspondentes e posteriormente fundidos por média. O conjunto resultante de features por voxel é passado a um VAE esparso, equipado com atenção de janelas deslocadas em 3D, que devolve a média e a variância dos vetores z_i. Três decodificadores paralelos são treinados contra a mesma latência: um rasterizador de Gaussians para fotorrealismo rápido, um decodificador de Radiance Fields baseado em decomposição para consistência multivista e um módulo FlexiCubes que extrai uma malha watertight via marching-cubes.

A geração propriamente dita segue um pipeline em duas fases regido por *Rectified Flow Transformers*, escolhidos por convergirem em menos passos que difusão. Na fase estrutural, o modelo sintetiza uma grade binária que indica quais voxels devem ser ativados. Para tornar o problema tratável, essa grade é antes compactada por um VAE 3D de baixa resolução. Na segunda fase, outro transformer, agora sensível à esparsidade, prediz os vetores z_i carregados nos voxels produzidos pelo estágio anterior. Ambos os fluxos empregam condicionamento via cross-attention: embeddings CLIP para texto e features DINO-v2 para imagens. Durante a amostragem, usa-se classifier-free guidance, com cerca de 50 passos totais, o que resulta em aproxidamente 10 s para Gaussians ou 18

s para malha em uma única GPU A100.

O treino utilizou aproximadamente 500 mil objetos oriundos de Objaverse-XL, ABO, 3D-FUTURE e HSSD, filtrados por licenças permissivas. Cada exemplar recebeu legendas automáticas via GPT-4o, e tanto texto quanto imagens foram submetidos a aumentos (variação de FOV, paráfrases e dropout de condição). Três escalas de modelo (342M, 1,1B e 2B de parâmetros) foram treinadas por 400.000 passos num cluster de 64 GPUs A100.

Os resultados experimentais demonstram que o SLAT reconstrói forma e textura com fidelidade superior a representações concorrentes como triplanes, latent point clouds ou vetores não estruturados, superando-as em PSNR, LPIPS, Chamfer Distance e F-score. Adicionalmente, a separação entre estrutura e detalhe habilita edições regionais: basta fixar a estrutura e regenerar apenas os latentes locais ou, inversamente, restringir o repaint a um subconjunto de voxels, preservando o restante do modelo.

Em síntese, o Trellis consolida uma estratégia de geração 3D que alia: (i) uma representação unificada, esparsa e editável; (ii) *Rectified Flow Transformers* para acelerar a amostragem; (iii) treinamento multiformato e *fitting-free*; e (iv) escalabilidade comprovada até bilhões de parâmetros sem requisito de memória cúbica na resolução.

3.3.2 Unique3D

O Unique3D é um sistema de geração que transforma uma única imagem em uma malha densa texturizada em poucos segundos, conciliando resolução, coerência multivista e eficiência computacional. Seu fluxo operacional inicia com um modelo de difusão multivista, ajustado a partir do modelo Stable Diffusion Image Variation, que sintetiza simultaneamente quatro vistas ortográficas (frontal, dorsal, superior e inferior) de 256px. Para garantir alinhamento geométrico, cada câmara é codificada por um índice inteiro injetado nas camadas de cross-attention (atenção cruzada). Essa codificação conjunta, por sua vez, faz com que o gerador aprenda dependências cruzadas entre as vistas e elimine artefatos de inconsistência comuns em abordagens que processam perspectivas isoladas.

Como a resolução inicial não captura detalhes finos, o Unique3D aplica um escalonamento progressivo. Primeiro, um modelo ControlNet multivista eleva as quatro imagens para 512, preservando a correspondência epipolar. Em seguida, o modelo Real-ESRGAN executa um redimensionamento, produzindo saídas de 2048. A mesma estratégia é aplicada, em paralelo, aos mapas de normais, assegurando que cor e geometria compartilhem granularidade espacial idêntica.

Para inferir relevos com precisão, o pipeline treina um modelo de difusão de normais cuja arquitetura replica a do gerador RGB, mas incorpora um U-Net de referência que guia o U-Net principal em nível pixel a pixel, reforçando o acoplamento entre

textura e geometria. Durante o treinamento, introduz-se um canal de *noise offset* que reduz o desvio estatístico entre ruído sintético e ruído gaussiano real, o que se reflete em superfícies mais suaves na inferência.

Concluída a síntese das vistas coloridas e de seus mapas de normais, entra em ação o ISOMER, cujo custo cresce linearmente com o número de faces. O algoritmo inicia com uma malha grossa estimada diretamente a partir dos mapas de normais e, depois, alterna passos de simplificação e refinamento que projetam gradientes fotométricos sobre faces mal alinhadas, enquanto um termo de expansão evita colapso em regiões planas. Esse procedimento gera malhas com dezenas de milhões de triângulos, compatíveis com motores como Blender e Unity, em menos de dez segundos em uma GPU A100.

Os dois modelos de difusão foram ajustados sobre um conjunto de aproximadamente 800.000 objetos: uma versão filtrada do Objaverse, complementada por Google Scanned Objects e ShapeNet, todos renderizados em quatro vistas ortográficas (RGB + normais) com fundo homogêneo. O treinamento consumiu quatro dias em oito GPUs RTX 4090, usando AdamW e batch-size efetivo de 192 na resolução de 256 px, reduzido pela metade nos níveis subsequentes devido às restrições de memória. O ISOMER não requer aprendizado.

Nos experimentos relatados pelos autores, o Unique3D supera Wonder3D, InstantMesh e OpenLRM em métricas objetivas, como CLIP-Score, LPIPS, Chamfer-Distance F_1, e em estudos de preferência humana, mantendo tempo de inferência de cerca de trinta segundos do clique inicial à malha final texturizada. A combinação de difusão multivista acoplada, *upscale* progressivo e reconstrução ISOMER comprova ser possível atingir alta fidelidade e consistência sem recorrer a representações volumétricas custosas ou otimizações caso a caso, razão pela qual o Unique3D se mostra adequado às exigências dessa aplicação.

3.3.3 Hunyuan3D 2.1

O Hunyuan3D 2.1 [6] é um sistema aberto de geração de ativos 3D que separa explicitamente a síntese de forma da síntese de aparência, estendendo a arquitetura do Hunyuan3D 2.0 com suporte a materiais fisicamente baseados (PBR) prontos para produção em motores gráficos como Unity e Unreal.

Na etapa de geometria, o modelo emprega um ShapeVAE que comprime a superfície do objeto em 1024 tokens latentes por meio de amostragem de importância em bordas e cantos, preservando detalhes geométricos finos. Sobre esse espaço comprimido opera o Hunyuan3D-DiT, um modelo de difusão baseado em Transformer com *Flow Matching*, treinado para mapear ruído gaussiano a latentes reconstruídos com alinhamento fino à imagem de entrada. O treinamento utilizou mais de um milhão de malhas provenien-

tes do Objaverse e Objaverse-XL, renderizadas sob 44 vistas predefinidas para maximizar a diversidade angular.

Na etapa de texturização, o módulo Hunyuan3D-Paint recebe mapas de posição e normais da malha gerada e produz materiais PBR de alta resolução. O pipeline gera seis imagens condicionadas por um backbone de difusão multivista, aplica um esquema de *view dropout* para garantir coerência 3D durante a inferência densa e empacota os resultados no espaço UV. Lacunas residuais são preenchidas por *inpainting* via projeção em vértices.

Em comparação aos demais modelos avaliados neste trabalho, o Hunyuan3D 2.1 apresenta o melhor equilíbrio entre fidelidade geométrica, qualidade de texturização e tempo de inferência para o cenário de entrada única, o que motivou sua adoção como modelo de reconstrução na etapa final do pipeline, após comparação qualitativa com o Trellis e o Unique3D.

3.4 Montagem do Museu Virtual

Após a geração dos ativos 3D, os objetos reconstruídos são importados para uma cena interativa desenvolvida em Unity, compondo um museu virtual baseado no acervo do Museu Egípcio do Cairo (MEC). Dentre os artefatos segmentados e reconstruídos, nove foram selecionados para compor a cena final, com base em dois critérios: (1) qualidade satisfatória da malha e da textura geradas pelo Hunyuan3D 2.1 e (2) disponibilidade de informações catalográficas suficientes para alimentar a base de conhecimento do assistente RAG, descrito na Seção 3.5. As malhas texturizadas geradas pelo modelo de reconstrução são exportadas em formato compatível com o Unity e importadas diretamente na cena.

Os artefatos reconstruídos são posicionados em um ambiente de galeria navegável, onde o usuário se locomove livremente e interage com os objetos por meio de um ponteiro de *raycast*, operado com o headset Meta Quest 2. Ao apontar para um artefato e acionar o controle, um painel é exibido na cena com informações contextuais sobre o objeto selecionado, geradas pelo assistente descrito na Seção seguinte.

Objetos que apresentaram desafios significativos de segmentação, como artefatos posicionados atrás de vitrines com reflexos especulares ou em condições de baixa luminosidade, não foram incluídos na cena final. Esses casos são discutidos na seção de resultados como limitações do pipeline proposto.

3.5 Assistente RAG

Para enriquecer a experiência do visitante com informações contextuais e verificáveis sobre os artefatos, integrou-se ao museu virtual uma camada de geração aumentada

Figura 3.4: *Cena VR interativa construída em Unity com artefatos reconstruídos do Museu Egípcio do Cairo. Ao apontar para um objeto com o ponteiro de raycast, um painel exibe a resposta gerada pelo assistente RAG.*

por recuperação (RAG, do inglês *Retrieval-Augmented Generation*). Essa abordagem ancora as respostas do modelo de linguagem em fontes documentais curadas, reduzindo o risco de alucinações e preservando a proveniência das informações apresentadas ao usuário [4].

3.5.1 Base de conhecimento e representação vetorial

A base de conhecimento foi construída a partir de informações catalográficas extraídas do site oficial do MEC [31]. Cada artefato é representado como um documento único, obtido pela concatenação dos seguintes campos: identificador, título, cultura, período, dinastia, datação, local de achado, descrições curta e longa, e palavras-chave. Optou-se por não aplicar fragmentação (*chunking*), uma vez que cada documento curatorial é relativamente curto e a fragmentação poderia separar informações semanticamente dependentes, como a associação entre período histórico e local de achado. O corpus resultante contém 46 documentos, correspondendo ao acervo selecionado do MEC.

Para a indexação vetorial, empregou-se o FAISS [8] com índice *IndexFlatIP* (*Inner Product*), operando como similaridade cosseno após normalização L2 dos vetores. A escolha do FAISS justifica-se pela simplicidade e eficiência para um dataset de pequeno porte que cabe inteiramente em memória. Os embeddings são gerados pelo modelo *text-embedding-3-small* da OpenAI, que produz vetores de 1536 dimensões. Esse modelo foi selecionado pelo equilíbrio entre custo computacional e qualidade de representação semântica, sendo suficiente para descrições curatoriais de extensão moderada.

3.5.2 Fluxo de recuperação e geração

O fluxo de interação opera da seguinte forma. Ao apontar para um artefato na cena Unity, o *raycast* retorna o identificador único do objeto selecionado. O cliente envia uma requisição GET ao *endpoint* REST `/query`, implementado em FastAPI (Python 3.11), contendo o identificador e a pergunta formulada pelo usuário. No servidor, o identificador seleciona o documento correspondente ao artefato, que é consultado no índice FAISS. Recuperam-se as $K = 2$ passagens mais similares à consulta, com um limiar mínimo de similaridade de 0,5; caso nenhum documento atinja esse limiar, o sistema retorna uma resposta genérica de indisponibilidade de informações. Não se emprega etapa de re-ordenação (*reranking*), dado que a busca já é restrita ao artefato identificado e o corpus é reduzido.

As passagens recuperadas são fornecidas como contexto ao modelo GPT-4o mini [17], selecionado por três razões: (1) baixa latência, compatível com a interação em tempo real requerida pela experiência VR; (2) custo reduzido por token, viável para um protótipo com múltiplas consultas por sessão; e (3) janela de contexto suficiente para respostas curtas fundamentadas em passagens recuperadas. O *prompt* do sistema configura o modelo como guia de museu, instruindo-o a utilizar exclusivamente o contexto fornecido, gerar respostas concisas de 20 a 40 palavras e produzir saída em JSON estruturado com três campos: texto da fala, pergunta de acompanhamento e fatos opcionais (1 a 3 itens). As respostas são geradas em língua inglesa.

3.6 Abordagem Proposta

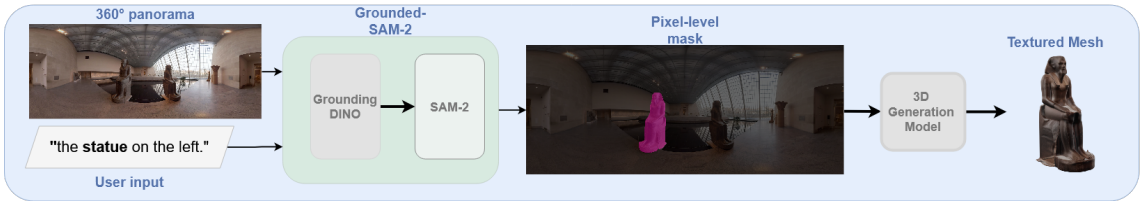

Figura 3.5: *Visão geral da etapa de geração 3D do pipeline proposto. Um panorama 360° e um prompt em linguagem natural são fornecidos ao módulo Grounded-SAM-2, que integra o Grounding DINO e o SAM 2 para produzir uma máscara de segmentação a nível de pixel. O objeto segmentado é então encaminhado ao modelo de geração 3D para síntese da malha texturizada.*

O pipeline proposto integra os componentes descritos nas seções anteriores em uma cadeia automatizada de ponta a ponta. Dada uma única imagem panorâmica equiretangular e um prompt em linguagem natural, o sistema produz uma malha 3D texturizada do artefato de interesse, que é posteriormente importada para o museu virtual

descrito na Seção 3.4 e associada à base de conhecimento do assistente RAG descrito na Seção 3.5.

A etapa de geração 3D, ilustrada na Figura 3.5, inicia com o método *Grounded-SAM-2* [7], que integra o Grounding DINO 1.5 [13] para detecção condicionada por texto e o SAM 2 [22] para refinamento de máscaras. Ambos os componentes aceitam entradas de grande campo de visão, eliminando a necessidade de dividir a imagem em blocos. Assim, o panorama completo de $8{,}192 \times 4{,}096$ pixels é processado em um único passo, com tempo de inferência de aproximadamente 0,5 s e consumo de cerca de 9GB de memória de GPU.

Máscaras com precisão de pixel são obtidas em uma única inferência. O recorte resultante é fornecido a um dentre três modelos avaliados: Trellis, Hunyuan3D 2.1 ou Unique3D, cada qual sintetizando uma malha 3D texturizada de alta qualidade. Com base na comparação qualitativa apresentada no Capítulo 4, o Hunyuan3D 2.1 foi adotado como modelo de reconstrução para a composição do museu virtual. Esse fluxo de trabalho elimina a captura multivisão, a segmentação manual e o pós-processamento, reduzindo o tempo de processamento por artefato para pouco mais de 60 s em uma única NVIDIA RTX 4090.

3.6.1 Avaliação com Usuários

Conduziu-se um estudo formativo de sessão única para avaliar se o pipeline VR proposto (i) suporta a interação com artefatos, (ii) permite a compreensão das informações entregues pelo assistente RAG e (iii) oferece uma experiência imersiva e utilizável para a recuperação de explicações por objeto.

O estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de Goiás (CEP/UFG) sob o protocolo CAAE nº 88442725.6.0000.5083, em 25 de julho de 2025, em conformidade com a Resolução CNS 466/2012. Todos os participantes assinaram o Termo de Consentimento Livre e Esclarecido (TCLE) antes da sessão experimental, sendo informados sobre os objetivos da pesquisa, procedimentos, riscos mínimos associados ao uso de headset VR, direito de desistência a qualquer momento e garantia de anonimato no tratamento dos dados.

A experiência foi conduzida em um headset Meta Quest 2, com sete participantes, todos estudantes de graduação em ciência da computação ou área afim, com idades entre 18 e 28 anos. Não foi estabelecido limite de tempo para a sessão. Cada participante foi instruído a interagir com todos os nove artefatos presentes na cena virtual e a ler o conteúdo informacional retornado pelo assistente RAG, exibido em painel textual na cena. Após a exploração, os participantes responderam a um questionário de oito itens (Q1–Q8), descritos na Tabela 3.1, abrangendo recordação de artefatos, retenção de co-

nhecimento via RAG, imersão, realismo percebido e usabilidade da interação por *raycast*. Para a verificação de recordação (Q1), a lista continha três artefatos presentes na cena dentre os nove disponíveis e dois distratores ausentes.

Tabela 3.1: *Questionário pós-experiência utilizado na avaliação com usuários.*

ID	Pergunta	Formato
Q1	Com quais artefatos você interagiu? (lista com 3 alvos presentes na cena e 2 distratores ausentes)	Seleção múltipla
Q2	Para o artefato selecionado, qual afirmação é correta? (recordação factual do conteúdo RAG)	Múltipla escolha
Q3	As informações sobre o artefato foram suficientes para a compreensão?	Sim/Não
Q4	Como você avalia seu senso de imersão no ambiente virtual?	Likert 1–5
Q5	Você percebeu o ambiente como realista?	Likert 1–5
Q6	Houve momentos em que você esqueceu que estava em um ambiente virtual?	Sim/Não
Q7	Quão fácil foi apontar para os artefatos e recuperar informações?	Likert 1–5
Q8	Quão aceitável foi o esforço necessário para apontar e recuperar informações?	Likert 1–5

CAPÍTULO 4

Resultados

Conforme discutido na Seção 3.3, a comparação entre os três modelos de reconstrução 3D adota uma abordagem qualitativa centrada na percepção visual, uma vez que não existem modelos 3D de referência (*ground truth*) dos artefatos museológicos avaliados e o objetivo final é a composição de uma experiência interativa em realidade virtual. Os critérios de avaliação incluem fidelidade das silhuetas em relação à imagem de entrada, qualidade percebida das texturas, hermeticidade da malha (ausência de buracos e auto-interseções) e preservação de detalhes geométricos finos. A Figura 4.1 compara Trellis, Hunyuan3D 2.1 e Unique3D lado a lado; cada malha é renderizada sob o mesmo ambiente HDRI neutro para garantir condições visuais equivalentes. Em seguida, apresenta-se um estudo de caso de ponta a ponta na Figura 3.3, medições de tempo de execução na Tabela 4.1 e resultados inconsistentes na Figura 4.2.

Tabela 4.1: *Tempo médio para gerar cada representação 3D.*

Modelos	Tempo médio
Trellis	8 s
Hunyuan3D 2.1	60 s
Unique3D	24 s

Para a análise qualitativa, utilizou-se o conjunto de dados extraído de vídeos em 4K de tours por três museus: Egyptian Museum in Cairo, Metropolitan Museum of Art em Nova York e National Museum of Natural History nas Filipinas. Quatro objetos da coleção foram escolhidos com base em sua qualidade e na iluminação global. Todas as malhas 3D foram geradas com uma única NVIDIA RTX 4090.

A partir da imagem 360° bruta (inferior) na Figura 3.3, o Grounding DINO localiza o artefato consultado ao desenhar uma caixa delimitadora condicionada ao texto (verde). O SAM 2 refina essa caixa em uma máscara com precisão de pixel (magenta) isolando a estátua do salão ao redor. O recorte mascarado é encaminhado ao modelo de reconstrução. O item no topo exibe quatro vistas da malha gerada pelo Hunyuan3D 2.1 (frontal, três-quartos à esquerda, três-quartos à direita e posterior). Esse exemplo demonstra o pipeline completo sem intervenção manual ou captura multivisão.

Figura 4.1: *Representações 3D geradas utilizando os modelos Trellis, Hunyuan3D 2.1 e Unique3D.*

Os quatro artefatos ilustrados na Figura 4.1 mostram um compromisso consistente na qualidade de saída entre os três modelos generativos. O Unique3D produz malhas cujas silhuetas e superfícies mais se aproximam das fotografias de entrada. Contudo, sua geometria frequentemente apresenta ruído de alta frequência e buracos ocasionais (por exemplo, fragmentos ausentes na bochecha do crânio e perfurações na placa). Trellis e Hunyuan3D 2.1 geram geometria mais limpa e hermética: os arcos dentários do crânio e a base do relevo estão totalmente fechados, sem auto-interseções visíveis. Essa suavidade sacrifica detalhes locais, pois alguns objetos perdem finos relevos para obter melhor qualidade global. Embora os modelos avaliados sejam capazes de produzir malhas e texturas de alta qualidade, os resultados também dependem da qualidade da entrada.

A Tabela 4.1 apresenta o tempo médio de parede (*wall-clock*) para geração da malha texturizada em uma NVIDIA RTX 4090, medido sobre os artefatos avaliados. A diferença expressiva entre os tempos reflete as distinções arquiteturais dos modelos: o Trellis emprega *Rectified Flow* com menos passos de amostragem, resultando no menor tempo de inferência (8 s); o Unique3D aplica *upscaling* progressivo em múltiplas resoluções (256, 512 e 2048 px), totalizando 24 s; o Hunyuan3D 2.1 inclui, além da geração geométrica, uma etapa separada de texturização PBR com seis imagens condicionadas e *inpainting* de lacunas, o que eleva o tempo total para aproximadamente 60 s. Considerando o conjunto dos critérios qualitativos e o tempo de inferência, a seção

seguinte descreve como o modelo selecionado foi empregado na composição do museu virtual e apresenta os resultados da avaliação com usuários.

4.1 Museu Virtual e Avaliação com Usuários

Com base nos resultados da comparação qualitativa entre os modelos, o Hunyuan3D 2.1 foi adotado como modelo de reconstrução para a composição do museu virtual, por apresentar o melhor equilíbrio entre fidelidade geométrica, qualidade de texturização e tempo de inferência. Os artefatos gerados foram importados para a cena Unity e associados às respectivas bases documentais do assistente RAG, conforme descrito nas Seções 3.4 e 3.5.

Para verificar a recordação dos artefatos (Q1), utilizou-se uma lista de imagens com cinco itens: três artefatos presentes na cena VR e dois distratores ausentes (Vaso Egípcio Antigo e Sarcófago de Tutancâmon). A Máscara Funerária de Tuya foi selecionada por todos os participantes (100%), seguida pelo Colosso de Amenhotep IV/Aquenáton (71%) e pela Cadeira da Princesa Satamun (57%). Apenas um participante selecionou um distrator (14%). Para a verificação factual (Q2), 50% dos participantes assinalaram a resposta correta, indicando que há espaço para aprimoramento na retenção factual. Esse resultado pode ser atribuído a múltiplos fatores: a complexidade do conteúdo histórico apresentado (nomes de faraós, dinastias e datas), o formato exclusivamente textual da informação exibida pelo painel RAG e a ausência de mecanismos de reforço, como repetição ou destaque visual dos dados mais relevantes. Além disso, a avaliação do assistente RAG neste estudo é limitada a uma medida indireta de retenção, sem métricas de qualidade de recuperação (relevância, completude ou posição do trecho correto no ranking).

Todos os participantes reportaram que as informações sobre os artefatos foram suficientes para a compreensão (Q3: 100% Sim). A aparente divergência entre a baixa retenção factual (Q2) e a percepção unânime de suficiência (Q3) sugere que os participantes compreenderam o conteúdo em nível geral durante a interação, mas a memorização de dados específicos demanda processamento cognitivo mais profundo, que os fatores mencionados acima não favoreceram. Para imersão (Q4), 42,9% avaliaram 3/5, 42,9% avaliaram 4/5 e 14,3% avaliaram 5/5. O realismo percebido (Q5) foi avaliado em 4/5 por 57,1% dos participantes. Apenas 28,6% reportaram momentos em que esqueceram estar em um ambiente virtual (Q6: Sim), enquanto 71,4% não relataram essa experiência. Esse resultado é consistente com as pontuações moderadas de imersão (Q4) e sugere que, embora o ambiente tenha sido percebido como envolvente, a ausência de estímulos complementares, como *feedback* háptico e áudio espacial, manteve a consciência metacognitiva dos participantes sobre a natureza virtual da experiência. Quanto à usabilidade da interação

por apontamento (Q7; 1 = muito fácil, 5 = muito difícil), 42,9% avaliaram 2/5, indicando baixa dificuldade percebida. O esforço requerido foi considerado majoritariamente aceitável (Q8): 42,9% avaliaram 4/5 e 28,6% avaliaram 5/5.

O tamanho da amostra é reduzido e o estudo não contou com um grupo de controle. Ainda assim, os resultados fornecem evidências iniciais de que os usuários conseguem identificar os artefatos com baixa taxa de falso reconhecimento e percebem as explicações do RAG como suficientes, motivando um estudo controlado de maior escala com instrumentos padronizados de presença e usabilidade.

4.1.1 Limitações

Embora o Hunyuan3D 2.1 seja capaz de produzir reconstruções de alta qualidade, os resultados são condicionados pelo sinal presente no recorte mascarado. Quando o recorte apresenta subexposição ou detalhes insuficientes, o gerador tende a suavizar excessivamente a geometria e a sintetizar texturas não fiéis à entrada, produzindo resultados de baixa fidelidade, conforme ilustrado na Figura 4.2. Esse comportamento evidencia a sensibilidade dos métodos de visão única às condições de iluminação e à distância de captura.

Figura 4.2: *Exemplo de resultado inconsistente gerado a partir de um recorte em ambiente de baixa luminosidade. O SAM 2 isola corretamente o objeto, mas a subexposição e a falta de detalhes levam o gerador a suavizar a geometria e a alucinar texturas.*

A segmentação de artefatos posicionados atrás de vitrines representa um desafio adicional: reflexos especulares, reflexos duplos, brilho e transparência parcial confundem o detector e o decodificador de máscaras, resultando em contornos fragmentados ou extrapolação para o fundo, como exemplificado na Figura 4.3. Esses artefatos contaminam a etapa de reconstrução, reduzindo a confiabilidade da malha e a fidelidade da textura.

Outros modos de falha incluem oclusões e aglomeração de objetos que causam máscaras incompletas e geometria ausente, estruturas pequenas ou finas que são ignoradas pelo detector ou erodidas pelo refinamento da máscara, e distorções equiretangulares próximas às costuras ou polos do panorama, onde o estiramento e as descontinuidades degradam tanto a detecção quanto as informações de reconstrução.

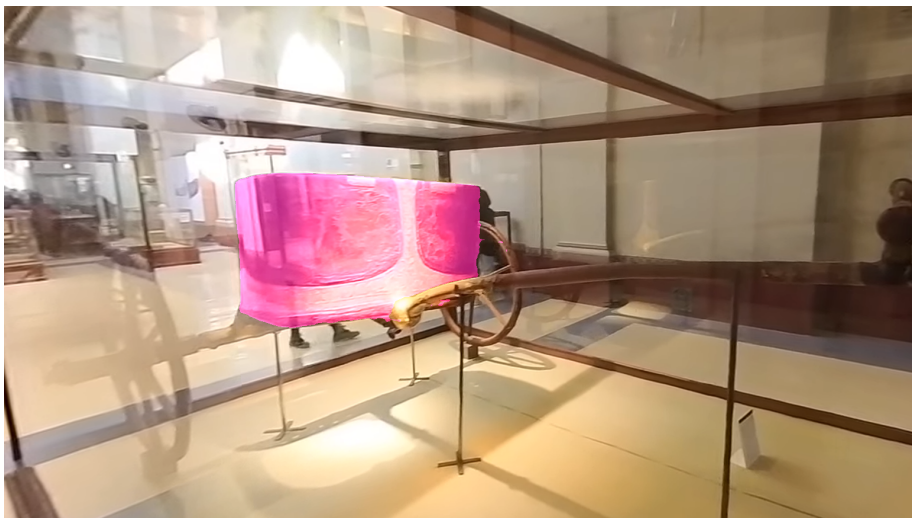

Figura 4.3: *Vista renderizada ilustrando como reflexos provenientes de vitrines de vidro resultam em máscaras de segmentação imprecisas, comprometendo a etapa de reconstrução 3D.*

CAPÍTULO 5

Conclusão

Este trabalho demonstra o potencial de modelos guiados por texto para gerar objetos 3D a partir de um único panorama 360°, destacando sua capacidade de extrair e reconstruir artefatos complexos do mundo real. Ao possibilitar a criação de representações 3D precisas com entrada mínima, esses modelos abrem novas perspectivas para aprimorar visitas a museus por meio de exibições interativas, enriquecendo o engajamento do público e a acessibilidade em experiências culturais virtuais. Nesse contexto, adotou-se uma abordagem em três etapas: (1) detecção e segmentação de objetos baseada em prompts, que isola artefatos de cenas 360° complexas a partir de um prompt textual; (2) reconstrução de objetos 3D, na qual cada artefato segmentado é transformado em um ativo 3D texturizado a partir de um único recorte mascarado; e (3) composição de um museu virtual interativo em Unity, enriquecido por um assistente de geração aumentada por recuperação (RAG) capaz de responder perguntas dos visitantes em inglês com base em fontes curatoriais verificáveis.

A comparação qualitativa entre Trellis, Unique3D e Hunyuan3D 2.1 evidenciou que o Hunyuan3D 2.1 oferece o melhor equilíbrio entre fidelidade geométrica, qualidade de texturização e tempo de inferência para o cenário de entrada única, motivando sua adoção como modelo de reconstrução na composição do museu virtual. A avaliação preliminar com usuários ($N = 7$) indicou que os participantes conseguiram interagir com os artefatos e recuperar explicações contextuais com esforço aceitável, reportando imersão moderada a forte e percebendo as informações fornecidas pelo assistente RAG como suficientes para a compreensão dos objetos.

Não obstante, os resultados evidenciaram que o pipeline é sensível à qualidade do sinal de entrada. Recortes provenientes de ambientes com subexposição ou detalhes insuficientes levaram o modelo de reconstrução a suavizar geometrias e sintetizar texturas pouco fiéis ao artefato original. De forma análoga, a segmentação de objetos posicionados atrás de vitrines com reflexos especulares produziu máscaras fragmentadas que comprometeram a etapa de reconstrução, conforme detalhado na Seção 4.1.1. Essas limitações restringem a aplicabilidade imediata do pipeline a cenários com condições mínimas de iluminação e visibilidade, e orientam diretamente as direções de trabalho futuro a seguir.

Tours virtuais podem ser enriquecidos com exibições reconstruídas apresentadas em galerias de VR imersivas ou sobrepostas como âncoras de AR no espaço físico do museu, permitindo que visitantes explorem artefatos frágeis em escala real, ajustem condições de iluminação e revelem detalhes ocultos que, de outra forma, permaneceriam inacessíveis.

Como trabalho futuro, planeja-se validar a abordagem em cenários mais diversos, variando layouts de galeria, tipos de objeto, condições de iluminação e disponibilidade de dados, a fim de avaliar robustez e capacidade de generalização. Pretende-se também incorporar um módulo de *relighting* para lidar com cenários de baixa luminosidade, aprimorar a robustez da segmentação em artefatos posicionados atrás de vitrines e quantificar as taxas de sucesso do pipeline de ponta a ponta. Por fim, estudos controlados com curadores e visitantes serão conduzidos para medir como esses ativos afetam a percepção de realismo e imersão, assegurando que os avanços técnicos se convertam em melhores experiências de VR e AR.

Referências

[1] BORTOLON, M.; TSESMELIS, T.; JAMES, S.; POIESI, F.; DEL BUE, A. **6dgs: 6d pose estimation from a single image and a 3d gaussian splatting model**. *arXiv preprint arXiv:2407.15484*, 2024.

[2] CARVAJAL, D. A. L.; MORITA, M. M.; BILMES, G. M. **Virtual museums. captured reality and 3d modeling**. *Journal of Cultural Heritage*, 45:234–239, 2020.

[3] CHEN, L.-C.; ZHU, Y.; PAPANDREOU, G.; SCHROFF, F.; ADAM, H. **Encoder–decoder with atrous separable convolution for semantic image segmentation**. *CoRR*, abs/1802.02611v3, 2018. arXiv preprint.

[4] COSSATIN, A. G.; MAURO, N.; FERRERO, F.; ARDISSONO, L. **Tell me more: integrating LLMs in a cultural heritage website for advanced information exploration support**. *Information Technology & Tourism*, 27:385–416, 2025.

[5] HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R. B. **Mask r-cnn**. *CoRR*, abs/1703.06870, 2017. arXiv preprint.

[6] HUNYUAN3D, T.; YANG, S.; YANG, M.; FENG, Y.; HUANG, X.; ZHANG, S.; OTHERS. **Hunyuan3D 2.1: From images to high-fidelity 3D assets with production-ready PBR material**. *arXiv preprint arXiv:2506.15442*, 2025.

[7] IDEA RESEARCH. **Grounded-sam-2: Promptable open-vocabulary segmentation**. https://github.com/IDEA-Research/Grounded-SAM-2, 2025. commit 2111d9c52c354671069dbd1ffe3acc09ce17068c, acesso em 21 jun. 2025.

[8] JOHNSON, J.; DOUZE, M.; JÉGOU, H. **Billion-scale similarity search with GPUs**. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

[9] KERBL, B.; KOPANAS, G.; LEIMKÜHLER, T.; DRETTAKIS, G. **3d gaussian splatting for real-time radiance field rendering.** *ACM Trans. Graph.*, 42(4):139–1, 2023.

[10] LAU, K. H. C.; BOZKIR, E.; GAO, H.; KASNECI, E. **Evaluating usability and engagement of large language models in virtual reality for traditional Scottish curling**. 2024. arXiv preprint arXiv:2408.09285.

[11] LEPOURAS, G.; KATIFORI, A.; VASSILAKIS, C.; CHARITOS, D. **Real exhibitions in a virtual museum**. *Virtual Reality*, 7:120–128, 2004.

[12] LIU, R.; WU, R.; HOORICK, B. V.; TOKMAKOV, P.; ZAKHAROV, S.; VONDRICK, C. **Zero-1-to-3: Zero-shot one image to 3d object**, 2023.

[13] LIU, S.; ZENG, Z.; REN, T.; LI, F.; ZHANG, H.; YANG, J.; LI, C.; GAO, J.; WANG, L.; LIU, Z.; ZHANG, L. **Grounding dino: Marrying dino with grounded pre-training for open-set object detection**. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 16898–16909, 2023.

[14] MAGNANI, M.; GUTTORM, A.; MAGNANI, N. **Three-dimensional, community-based heritage management of indigenous museum collections: Archaeological ethnography, revitalization and repatriation at the sámi museum siida**. *Journal of Cultural Heritage*, 31:162–169, 2018.

[15] MILDENHALL, B.; SRINIVASAN, P. P.; TANCIK, M.; BARRON, J. T.; RAMAMOORTHI, R.; NG, R. **Nerf: Representing scenes as neural radiance fields for view synthesis**, 2020.

[16] NICCOLUCCI, F. **Virtual reality in archaeology: a useful tool or a dreadful toy?** *Mediaterra Art & Technology Festival*, 99, 1999.

[17] OPENAI. **GPT-4 technical report**. Technical report, OpenAI, 2023. arXiv preprint arXiv:2303.08774.

[18] ORR, M.; POITRAS, E.; BUTCHER, K. R. **Informal learning with extended reality environments: Current trends in museums, heritage, and tourism**. In: *Augmented reality in tourism, museums and heritage: A new technology to inform and entertain*, p. 3–26. Springer, 2021.

[19] PANTILE, D.; FRASCA, R.; MAZZEO, A.; VENTRELLA, M.; VERRESCHI, G. **New technologies and tools for immersive and engaging visitor experiences in museums: The evolution of the visit-actor in next-generation storytelling, through augmented and virtual reality, and immersive 3d projections**. In: *2016 12th International conference on signal-image technology & internet-based systems (SITIS)*, p. 463–467. IEEE, 2016.

[20] RABBY, A.; ZHANG, C. **Beyondpixels: A comprehensive review of the evolution of neural radiance fields**. *arXiv preprint arXiv:2306.03000*, 2023.

[21] RADL, L.; STEINER, M.; PARGER, M.; WEINRAUCH, A.; KERBL, B.; STEINBERGER, M. **Stopthepop: Sorted gaussian splatting for view-consistent real-time rendering**. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024.

[22] RAVI, N.; GABEUR, V.; HU, Y.-T.; HU, R.; RYALI, C.; MA, T.; KHEDR, H.; RÄDLE, R.; ROLLAND, C.; GUSTAFSON, L.; MINTUN, E.; PAN, J.; ALWALA, K. V.; CARION, N.; WU, C.-Y.; GIRSHICK, R.; DOLLÁR, P.; FEICHTENHOFER, C. **SAM 2: Segment anything in images and videos**. *arXiv preprint arXiv:2408.00714*, 2024. Versão 2 (revisado em 28 out. 2024).

[23] REDMON, J.; DIVVALA, S.; GIRSHICK, R.; FARHADI, A. **You only look once: Unified, real-time object detection**. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 779–788, 2016.

[24] REMONDINO, F. **Heritage recording and 3D modeling with photogrammetry and 3D scanning**. *Remote Sensing*, 3(6):1104–1138, 2011.

[25] ROUSSOU, M. **Immersive interactive virtual reality in the museum**. *Proc. of TiLE (Trends in Leisure Entertainment)*, 2001.

[26] S MEYER, L. S.; AAEN, J. E.; TRANBERG, A. R.; KUN, P.; FREIBERGER, M.; RISI, S.; LØVLIE, A. S. **Algorithmic ways of seeing: Using object detection to facilitate art exploration**. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, p. 1–18. Association for Computing Machinery, 2024.

[27] SCOPIGNO, R.; CALLIERI, M.; CIGNONI, P.; CORSINI, M.; DELLEPIANE, M.; PONCHIO, F.; RANZUGLIA, G. **3D models for cultural heritage: Beyond plain visualization**. *Computer*, 44(7):48–55, 2011.

[28] SHI, R.; CHEN, H.; ZHANG, Z.; LIU, M.; XU, C.; WEI, X.; CHEN, L.; ZENG, C.; SU, H. **Zero123++: a single image to consistent multi-view diffusion base model**. *arXiv preprint arXiv:2310.15110*, 2023.

[29] SITZMANN, V.; ZOLLHÖFER, M.; WETZSTEIN, G. **Scene representation networks: Continuous 3d-structure-aware neural scene representations**, 2020.

[30] SKAMANTZARI, M.; GEORGOPOULOS, A. **3d visualization for virtual museum development**. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41:961–968, 2016.

[31] THE EGYPTIAN MUSEUM IN CAIRO. **The egyptian museum in cairo (official website)**. https://egyptianmuseumcairo.eg/. Acessado em: 02 jan. 2026.

[32] TOCHILKIN, D.; PANKRATZ, D.; LIU, Z.; HUANG, Z.; LETTS, A.; LI, Y.; LIANG, D.; LAFORTE, C.; JAMPANI, V.; CAO, Y.-P. **Triposr: Fast 3d object reconstruction from a single image**. *arXiv preprint arXiv:2403.02151*, 2024.

[33] VASIC, I.; FILL, H.-G.; QUATTRINI, R.; PIERDICCA, R. **LLM-aided museum guide: Personalized tours based on user preferences**. In: *Extended Reality: International Conference, XR Salento 2024*, volume 15029 de **Lecture Notes in Computer Science**, p. 249–262. Springer Nature Switzerland, 2024.

[34] WANG, Z.; YUAN, L.; WANG, L.; JIANG, B.; ZENG, W. **VirtuWander: Enhancing multi-modal interaction for virtual tour guidance through large language models**. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.

[35] WU, K.; LIU, F.; CAI, Z.; YAN, R.; WANG, H.; HU, Y.; DUAN, Y.; MA, K. **Unique3d: High-quality and efficient 3d mesh generation from a single image**. *arXiv preprint arXiv:2405.20343*, 2024.

[36] WU, T.; YUAN, Y.-J.; ZHANG, L.-X.; YANG, J.; CAO, Y.-P.; YAN, L.-Q.; GAO, L. **Recent advances in 3d gaussian splatting**. *Computational Visual Media*, p. 1–30, 2024.

[37] XIANG, J.; LV, Z.; XU, S.; DENG, Y.; WANG, R.; ZHANG, B.; CHEN, D.; TONG, X.; YANG, J. **Structured 3d latents for scalable and versatile 3d generation**. *arXiv preprint arXiv:2412.01506*, 2024.

[38] XU, J.; CHENG, W.; GAO, Y.; WANG, X.; GAO, S.; SHAN, Y. **Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models**. *arXiv preprint arXiv:2404.07191*, 2024.

[39] YU, A.; YE, V.; TANCIK, M.; KANAZAWA, A. **pixelnerf: Neural radiance fields from one or few images**, 2021.