

SONA: Geração Automatizada de Audiodescrição com Ajuste Temporal a Lacunas de Fala

Pedro Alves Valentim
pedroalvesvalentim@midia.com.uff.br
Laboratório MídiaCom, Instituto de Computação
Universidade Federal Fluminense
Niterói, Rio de Janeiro, Brazil

Débora C. Muchaluat-Saade
debora@midia.com.uff.br
Laboratório MídiaCom, Instituto de Computação
Universidade Federal Fluminense
Niterói, Rio de Janeiro, Brazil

ABSTRACT

Audio description (AD) is a critical accessibility resource for blind and low-vision audiences, but producing AD scripts remains time-intensive and costly when performed manually. This paper introduces SONA (System for Optimized Narration and Audiodescription), a hybrid, end-to-end pipeline that generates AD assets directly from a final audiovisual media file with minimal user input. SONA segments a video into shots, extracts and clusters faces for character identification, tracks characters through color-based visual annotations, and combines environment summarization with structured inference of actions, interactions, facial expressions, and salient objects using multimodal large language models. The resulting per-shot narrations are orchestrated into speech-free intervals detected by voice activity detection, producing a timestamped WebVTT AD script and a synthesized AD track with configurable output language and voice. A core contribution is an explicit fit-to-gap mechanism that accepts a narration segment only when its measured synthesized duration respects the assigned silence gap, using bounded time compression and iterative LLM-based shortening when needed. We evaluate SONA on MPII-MD, TV-AD, and a full-length feature film case study, complement the analysis with component ablations, and report an expert qualitative assessment of two complete film outputs. Results indicate low lexical overlap but stronger semantic agreement, a clear timing benefit from fit-to-gap enforcement, and practical usefulness as a drafting tool for AD workflows.

KEYWORDS

Audiodescrição, acessibilidade, compreensão de vídeo, modelos de linguagem, alinhamento temporal, síntese de voz.

1 INTRODUÇÃO

Mídias audiovisuais são hoje um elemento central da cultura e do entretenimento, mas ainda impõem barreiras relevantes para pessoas cegas ou com baixa visão. A audiodescrição (AD) busca reduzir essa barreira por meio de uma faixa narrativa adicional que descreve informações visuais que não podem ser inferidas apenas pelo áudio original, como ações, cenário,

expressões faciais e elementos importantes da narrativa. Essa narração deve ser inserida nos intervalos sem fala, preservando diálogos, trilha sonora e efeitos sonoros.

A produção manual de roteiros de AD, entretanto, ainda é custosa e pouco escalável. Estima-se que um profissional leve, em média, cerca de 60 horas para produzir a audiodescrição de um filme de 2 horas [8]. Esse custo limita a disponibilidade do recurso, especialmente em produções menores, que não contam com o orçamento de grandes estúdios. Nesse contexto, soluções automatizadas podem reduzir tempo de produção, ampliar cobertura e servir como apoio ao fluxo de trabalho de audiodescritores.

Este trabalho propõe o SONA, *System for Optimized Narration and Audiodescription*, uma solução para geração automatizada de faixas de audiodescrição a partir do arquivo audiovisual final. O sistema combina segmentação em tomadas, extração e agrupamento de faces para identificação de personagens, detecção de intervalos de fala, descrição do ambiente físico e inferência de ações, objetos e interações entre personagens. Essas informações são organizadas por módulos que utilizam modelos de linguagem multimodais e LLMs para gerar descrições narrativas, posteriormente sincronizadas aos intervalos sem fala. Como saídas, o sistema produz um roteiro de AD em WebVTT [20], uma faixa de AD sintetizada e um vídeo de demonstração com a faixa mixada ao áudio original.

A principal contribuição técnica do SONA é o controle explícito de ajuste ao intervalo disponível: antes de agendar uma narração, o sistema sintetiza o áudio, mede sua duração e só aceita o segmento se ele couber no intervalo de silêncio detectado. Quando necessário, aplica aceleração limitada ou encurtamento iterativo por LLM. Avaliamos o sistema com métricas textuais em MPII-MD, TV-AD e em um estudo de caso de longa-metragem, complementando os resultados com ablações de componentes e uma avaliação qualitativa por especialista em audiodescrição.

2 TRABALHOS RELACIONADOS

CineAD [3] é um sistema de geração automática de roteiros de audiodescrição projetado especialmente para filmes brasileiros em português. Sua metodologia parte de artefatos textuais associados ao filme, em particular roteiro e legendas, para extrair ações, personagens e intervalos sem diálogo. A partir desses elementos, o sistema resume sentenças relevantes e as aloca em lacunas temporais disponíveis. Essa estratégia tem a vantagem de explorar informação narrativa já estruturada, mas também impõe uma restrição importante: o roteiro

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

precisa estar disponível e corresponder de forma precisa ao corte final do material audiovisual. Em produções nas quais o roteiro não existe, não está sincronizado ou foi alterado durante a edição, essa dependência reduz a aplicabilidade da abordagem.

AutoAD-Zero [22] propõe um *framework* zero-shot baseado em VLMs e LLMs, sem ajuste fino. O sistema primeiro gera descrições visuais detalhadas, com foco em personagens, ações, interações e expressões faciais, e depois condensa essas informações em sentenças de AD. Uma contribuição relevante dessa linha é mostrar que modelos multimodais de uso geral já podem produzir descrições razoáveis sem treinamento específico para audiodescrição. Por outro lado, o controle temporal permanece principalmente associado ao orçamento textual e à duração esperada, não a uma verificação final da duração efetivamente sintetizada.

Shot-by-Shot [21] segue uma direção próxima, mas incorpora tomadas como unidades básicas de compreensão do vídeo e usa pistas de gramática cinematográfica, como enquadramento, continuidade e relação entre tomadas adjacentes, para orientar a descrição. Essa escolha é particularmente relevante para AD, pois diferentes tipos de enquadramento favorecem diferentes prioridades descritivas: planos fechados tendem a destacar expressões faciais, enquanto planos abertos favorecem ambiente, posição espacial e movimento. MMAD [23], por sua vez, adota uma arquitetura multimodal treinada, combinando informação de áudio, marcação de atores e alinhamento contextual com cliques legendados. Essa abordagem explora sinais mais ricos, mas exige dados, treinamento e complexidade adicional.

Esses trabalhos diferem principalmente nos artefatos de entrada exigidos, na necessidade de treinamento, na forma de lidar com personagens e no controle temporal da narração. Em contraste com abordagens dependentes de roteiro, legenda ou treinamento específico, o SONA opera diretamente sobre a mídia audiovisual final. A proposta combina extração visual de identidade e eventos com geração textual por LLMs, produz tanto roteiro quanto faixa sintetizada, e verifica o ajuste temporal a partir da duração medida da fala gerada. Portanto, a diferença central não está apenas em gerar uma sentença de AD, mas em transformar essa sentença em um artefato temporizado e audível, respeitando as lacunas de fala detectadas no áudio original. A Tabela 1 resume essas diferenças.

3 ARQUITETURA

A única entrada do sistema é o arquivo de vídeo da obra, em formatos como MP4 ou MKV. A partir dele, o SONA executa três procedimentos iniciais: extração de intervalos sem fala, segmentação em planos e extração de personagens. Antes da execução, o usuário pode configurar hiperparâmetros globais, incluindo idioma de saída e voz usada na síntese final. Durante a execução, a única interação humana é a atribuição opcional de nomes aos agrupamentos de personagens encontrados. Se o usuário não fornecer nomes, o sistema usa pronomes ou descrições visuais, como “o homem de terno” ou “a mulher de

vermelho”. Essa escolha mantém o sistema operável mesmo quando não há nomeação de personagens.

Para cada plano segmentado, o sistema executa uma subpipeline de descrição visual. Primeiro são geradas uma descrição do ambiente físico e uma versão anotada do plano com marcações coloridas sobre personagens reconhecidos. Em seguida, o sistema infere personagens principais, ações, interações, expressões faciais e objetos relevantes. Essas informações são condensadas em uma sentença de AD para o plano, que passa a ser a unidade textual usada pelo orquestrador. A divisão por planos busca evitar que mudanças abruptas de câmera sejam tratadas como continuidade visual, ao mesmo tempo em que preserva unidades suficientemente curtas para inferência multimodal.

A Figura 1 resume o fluxo. A extração de intervalos de fala por VAD (A), a segmentação em planos (B) e a extração de personagens (C) precedem o laço principal por plano: descrição de ambiente (D), marcação de personagens (E), reconhecimento de personagens (F), inferência de ações e interações (G) e composição da legenda narrativa (H). Ao final, as descrições são orquestradas (I) nos intervalos sem fala, opcionalmente traduzidas e sintetizadas, produzindo o roteiro WebVTT e a faixa de AD.

A arquitetura separa deliberadamente percepção, geração textual e controle temporal. Os módulos de visão computacional fornecem evidências estruturadas sobre planos e personagens, os modelos multimodais e LLMs transformam essas evidências em linguagem natural, e o orquestrador verifica se a narração gerada pode ser inserida no áudio original sem invadir fala detectada. Essa separação facilita substituições futuras de modelos, já que um detector de faces, VLM ou TTS mais robusto pode ser incorporado sem alterar o desenho geral da *pipeline*.

4 IMPLEMENTAÇÃO

O sistema foi implementado em Python [19]. OpenCV [2] é usado em tarefas de visão computacional ao longo da *pipeline*, enquanto FFmpeg [6] é usado para extração de áudio, segmentação de vídeo e geração de artefatos demonstrativos. O comportamento dos módulos é controlado por hiperparâmetros, como caminhos de saída, *prompts*, *seeds* dos modelos, idioma alvo e voz de TTS.

A extração de intervalos de fala (A) inicia pela separação da trilha de áudio do vídeo. Em seguida, o módulo aplica um modelo pré-treinado da biblioteca Silero VAD [18], com leitura padronizada do áudio a 16 kHz, para detectar segmentos de fala. Os intervalos de silêncio são obtidos como complemento temporal dos segmentos detectados e podem ser subdivididos por um limiar máximo configurável, evitando lacunas longas demais para a etapa de alocação. Cada lacuna é representada por tempos de início e fim e por uma duração conservadora. Na implementação atual, a duração usada nas verificações posteriores é arredondada para baixo, reduzindo a chance de que diferenças de arredondamento façam uma narração ultrapassar a fala original. O resultado é uma lista ordenada de intervalos disponíveis para inserção de AD.

Tabela 1: Comparação com trabalhos relacionados. O símbolo ~ indica suporte parcial ou ausência de garantia estrita.

Trabalho	Roteiro	Treinamento	Nomeação de Personagens	Ajuste Temporal	Saída
CineAD [3]	Sim	Não	Não	~	Texto
AutoAD-Zero [22]	Não	Não	Sim	~	Texto
MMAD [23]	Não	Sim	Sim	~	Texto
Shot-by-Shot [21]	Não	Não	~	~	Texto
SONA	Não	Não	~	Sim	WebVTT + faixa de AD

Trabalho	Método e limitação principal
CineAD [3]	Usa roteiro e legendas para resumir ações e alocá-las em intervalos sem diálogo; é sensível a desalinhamento entre roteiro, legendas e corte final.
AutoAD-Zero [22]	Usa VLMs e LLMs <i>off-the-shelf</i> para gerar descrições de personagens e ações; controla concisão por texto, mas não por duração sintetizada medida.
MMAD [23]	Integra áudio, marcação de atores e contexto legendado em uma arquitetura treinada; exige dados, treinamento e maior complexidade operacional.
Shot-by-Shot [21]	Usa tomadas e gramática cinematográfica para melhorar <i>grounding</i> e seleção de eventos; não impõe ajuste à lacuna de narração.
SONA	Combina visão computacional, VLMs/LLMs e TTS medido para produzir roteiro e faixa audível; depende da qualidade do VAD.

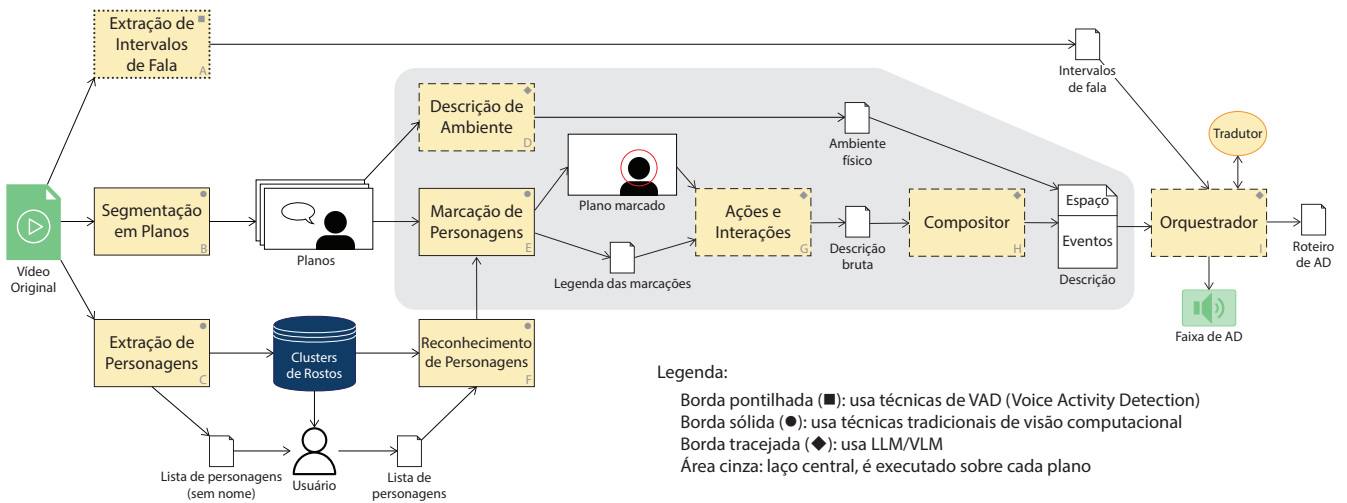


Figura 1: Pipeline do SONA.

Essa normalização dos intervalos é importante porque o restante da *pipeline* trabalha com lacunas discretas, não com a faixa contínua de áudio. Intervalos muito longos são subdivididos para evitar que uma única descrição seja posicionada em uma janela excessivamente ampla, o que poderia gerar deslocamento perceptual em relação ao evento visual descrito. Intervalos muito curtos são descartados posteriormente pelo orquestrador. Assim, o sistema separa duas decisões: primeiro identifica regiões potencialmente livres de fala, depois decide se uma sentença específica, sintetizada com uma voz específica, cabe em uma dessas regiões.

A segmentação em planos (B) usa PySceneDetect [4] para identificar mudanças visuais no conteúdo e exportar cada plano como um clipe separado. O módulo aplica detecção

adaptativa de conteúdo, converte os limites temporais para segundos e registra, para cada plano, seu início e fim na linha do tempo do vídeo original. Essa informação é usada posteriormente para associar descrições locais a regiões temporais do filme. A segmentação física dos cliques é realizada com FFmpeg, de modo que os módulos multimodais possam processar trechos curtos em vez do vídeo completo.

A extração de personagens (C) amostra quadros do vídeo, detecta faces e calcula *embeddings* de identidade com Insight-Face [5]. O ritmo de amostragem é configurável: pode-se usar todos os quadros ou reduzir a taxa para diminuir custo computacional. Para cada face detectada, o sistema armazena a caixa delimitadora e um *embedding* de 512 dimensões, normalizado para comparação por distância ou similaridade. Esses

embeddings são agrupados no espaço vetorial, por padrão com HDBSCAN [10], para identificar aparições recorrentes sem exigir que o número de personagens seja definido previamente. Para cada grupo não considerado ruído, o sistema seleciona um conjunto de exemplos representativos e permite que o usuário associe opcionalmente um nome ao personagem. O resultado é um conjunto de instâncias faciais com rótulos derivados dos agrupamentos, que pode ser reutilizado por módulos posteriores sem recalculá-los todas as vezes.

Cada plano é então processado pelos módulos de descrição de ambiente (D) e marcação de personagens (E). A descrição de ambiente usa um LLM multimodal local servido via Ollama [17], a partir de quadros amostrados do plano, para produzir uma frase curta no estilo de audiodescrição. Essa frase não tenta substituir a descrição de ação; seu papel é fornecer contexto espacial ou visual estável, como local, iluminação, disposição geral ou atmosfera da cena.

A marcação de personagens projeta os agrupamentos de identidade de volta sobre o vídeo. Em cada quadro do plano, faces detectadas são comparadas às identidades já rotuladas por similaridade de *embedding*. Quando há correspondência, a face recebe uma marcação colorida — consistente dentro do plano — e uma legenda num arquivo auxiliar associa cada personagem a uma cor. A paleta é fixa e limitada a sete cores nomeáveis, pois mais a frente uma VLM deve interpretar nomes de cores, não outras formas de representação (e.g. valores RGB). Essa estratégia segue evidências de que marcações visuais simples, como círculos coloridos, podem ser mais eficazes do que recortes de imagem [12, 16]. A decisão também evita inserir texto diretamente sobre o vídeo, o que poderia prejudicar a leitura visual do modelo ou criar artefatos ambíguos.

O módulo de ações e interações (G) recebe o plano anotado e a legenda de cores, quando disponível, e gera uma descrição estruturada dos elementos visuais relevantes: personagens principais, ações, interações, expressões faciais e objetos. Quando o usuário nomeia personagens, o *prompt* inclui o mapeamento entre nomes e cores; quando não há nomes, o *prompt* orienta o modelo a usar descrições visuais dos próprios personagens, por exemplo com base em roupa, posição ou aparência. A saída segue um molde estruturado, o que reduz variações indesejadas e facilita a etapa seguinte. Se o modelo não preencher campos relevantes, a composição narrativa do plano pode ser ignorada, evitando que sentenças vazias ou genéricas sejam encaminhadas ao orquestrador.

O módulo compositor (H) condensa a descrição estruturada em uma única sentença de AD por plano. Essa sentença deve ser curta, informativa e narrável, pois será posteriormente encaixada em intervalos de silêncio potencialmente pequenos. O objetivo não é produzir uma legenda exaustiva do vídeo, mas selecionar uma formulação compatível com convenções de AD: foco em informação visual relevante, linguagem direta e ausência de explicações interpretativas quando a imagem permite uma descrição mais neutra. A descrição de ambiente e a descrição de ações são combinadas em uma descrição-mestra do plano, usada como unidade textual pelo orquestrador.

O orquestrador (I) alinha as sentenças de AD aos intervalos sem fala detectados no início da *pipeline*. Primeiro, ele carrega as descrições produzidas pelo compositor, os limites temporais dos planos e a lista de lacunas geradas pelo VAD. Lacunas menores que uma duração mínima configurável são descartadas, pois tendem a produzir narrações curtas demais para serem úteis ou perceptualmente confortáveis. Para cada descrição candidata, o sistema busca um intervalo compatível com o plano correspondente e tenta inserir a narração sem sobrepor segmentos já agendados.

Após escolher um intervalo, o sistema aplica tradução opcional para o idioma alvo e sintetiza a narração com o modelo local Kokoro [7]. A duração do áudio sintetizado é medida com Pydub [14]. Essa medição é necessária porque o número de caracteres ou palavras não prediz de forma confiável a duração final da fala: voz, idioma, pontuação, prosódia e TTS afetam o tempo real. Se a fala couber no intervalo, o segmento é aceito. Caso contrário, o sistema calcula o fator de aceleração necessário. Se esse fator não ultrapassar o limite máximo configurado, a faixa é acelerada e aceita; se ultrapassar, a aceleração é rejeitada para preservar inteligibilidade.

Quando a aceleração aceitável não é suficiente, o sistema aciona um encurtamento por LLM. O encurtamento é aplicado à descrição-fonte, antes da tradução final, para preservar controle sobre o conteúdo original e evitar encurtar diretamente uma tradução que possa introduzir perdas adicionais. Em seguida, o texto encurtado é traduzido novamente quando necessário, sintetizado e medido. Esse ciclo de encurtamento, tradução, síntese e medição se repete até que a narração caiba no intervalo ou até que não haja uma alternativa válida para o plano. Na prática, isso faz com que o texto final seja determinado não apenas por concisão textual, mas pela duração audível da fala resultante.

Essa etapa impõe a principal garantia operacional do SONA: um segmento de AD só é agendado se sua duração sintetizada medida não exceder a duração do intervalo de silêncio atribuído. O tempo final no WebVTT é calculado a partir da duração efetiva do áudio aceito, mantendo consistência entre roteiro e faixa gerada. Se nenhuma lacuna compatível existir para um plano, a narração desse plano é omitida, em vez de ser forçada sobre diálogo. A garantia, naturalmente, depende da qualidade dos intervalos detectados por VAD: se houver fala residual dentro de um intervalo classificado como silêncio, ainda pode haver sobreposição perceptual apesar da restrição de duração.

Após aceitar os segmentos, o sistema sobrepõe cada clipe de AD em uma trilha silenciosa com a mesma linha do tempo do vídeo, exporta a faixa de AD e grava o roteiro WebVTT sincronizado. Por fim, uma chamada ao FFmpeg gera um vídeo de demonstração com o áudio original e a faixa de AD mixados. Esse arquivo não é tratado como produto final de distribuição, mas como artefato de validação, permitindo que o usuário ou avaliador assista rapidamente ao resultado temporizado.

5 AVALIAÇÃO

Avaliamos o SONA por duas perspectivas complementares. Primeiro, realizamos uma avaliação automática baseada em referência, comparando o texto gerado com descrições profissionais em três cenários: clipes do MPII-MD [15], segmentos do TV-AD [22] e um estudo de caso de longa-metragem usando *Night of the Living Dead*. Essa avaliação é complementada por uma ablação direcionada dos módulos de marcação de personagens e ajuste ao intervalo. Segundo, analisamos qualitativamente saídas completas de dois filmes, com avaliação de um especialista em audiodescrição. A combinação é importante porque métricas textuais ajudam a comparar com referências, mas não capturam integralmente requisitos de AD, como *timing*, preservação da trilha original e utilidade editorial.

5.1 Avaliação quantitativa

MPII-MD é um *benchmark* de descrição de vídeos derivado de faixas profissionais de AD alinhadas a intervalos de filmes. O conjunto contém mais de 68 mil pares sentença-clipe de 94 filmes, cobrindo grande diversidade de gêneros, estilos visuais e contextos narrativos. Essa diversidade torna o conjunto útil para avaliar generalização, mas também aumenta o desacordo esperado entre uma descrição gerada automaticamente e uma referência profissional: duas descrições podem estar visualmente corretas e ainda assim selecionar eventos diferentes.

TV-AD, por sua vez, reúne descrições profissionais de episódios de séries, com 34.013 instâncias de AD em 426 episódios. Seus intervalos de narração são curtos em média, com duração reportada de 2,07s, o que torna o conjunto particularmente relevante para avaliar descrições compactas sob restrições temporais. Diferentemente de um filme completo, muitos segmentos de TV-AD correspondem a unidades curtas e centradas em ação, próximas do regime em que o SONA tenta operar: identificar rapidamente quem aparece, o que muda visualmente e como expressar isso em uma frase narrável.

Para complementar os *benchmarks* de clipes, realizamos também um estudo de caso de longa-metragem. Nesse cenário, comparamos o roteiro completo gerado pelo SONA para *Night of the Living Dead* com um roteiro profissional da mesma obra, concatenando as sentenças em ordem temporal. Essa comparação não avalia diretamente se cada sentença foi dita no melhor instante possível, mas permite observar se o texto gerado mantém alinhamento semântico global com a referência ao longo de uma obra completa.

Para MPII-MD e TV-AD, cada instância do conjunto foi tratada como um intervalo de referência, e a saída do SONA para o segmento correspondente foi comparada com a descrição profissional. Para o longa-metragem, calculamos as mesmas métricas sobre o roteiro completo. Os resultados devem ser interpretados considerando o contexto de inferência utilizado: nos experimentos reportados, os módulos multimodais processam três quadros por plano e usam Qwen3-VL:8B como VLM, escolha motivada pelas restrições de uma GPU

NVIDIA RTX 4060 com 8 GB de VRAM, visando representar o uso do sistema numa máquina comum. Assim, os números reportados não devem ser lidos como limite superior da abordagem, mas como uma estimativa conservadora sob *hardware* de consumo.

Como a AD admite várias formulações válidas para o mesmo conteúdo visual, reportamos métricas de sobreposição lexical e similaridade semântica. BLEU [11], ROUGE [9] e METEOR [1] capturam diferentes formas de sobreposição de *tokens*. BERTScore [24] e Sentence-BERT [13], por outro lado, comparam representações semânticas e são mais tolerantes a paráfrases. Reportamos BERTScore F1 por equilibrar precisão e cobertura semântica: precisão isolada favoreceria saídas curtas que acertam apenas parte da referência, enquanto *recall* isolada favoreceria saídas mais longas que cobrem mais conteúdo, mas podem introduzir detalhes não sustentados. Ainda assim, métricas textuais não medem diretamente *grounding* visual, alocação temporal, ausência de alucinação ou utilidade narrativa para acessibilidade; por isso, elas são usadas aqui como *proxy* inicial, não como medida completa de qualidade de AD.

Tabela 2: Similaridade entre saídas do SONA e referências profissionais de AD. Reportamos métricas lexicais e semânticas. Para todas estas métricas maior é melhor.

Dataset	BLEU	METEOR	R-1	R-2	R-L
MPII-MD	0,38%	7,18%	5,97%	0,75%	5,28%
TV-AD	1,14%	16,74%	14,59%	1,82%	12,86%
<i>Night</i> (filme)	0,95%	13,52%	17,62%	1,96%	14,39%
Dataset	BERTScore (F1)		Sentence-BERT		
MPII-MD	45,04%		20,00%		
TV-AD	86,55%		37,26%		
<i>Night</i> (filme)	87,03%		30,82%		

A Tabela 2 mostra baixa sobreposição lexical em todos os cenários, especialmente em BLEU e ROUGE-2. Esse resultado era esperado em parte, pois descrições profissionais podem escolher eventos, granularidades e formulações diferentes para o mesmo trecho visual. Em MPII-MD, a diversidade de filmes, entidades e contextos de enredo dificulta ainda mais a correspondência lexical e reduz também os escores semânticos. Em termos práticos, os resultados indicam que o SONA captura parte do conteúdo visual relevante, mas frequentemente diverge da referência em seleção de detalhes ou em formulação textual.

Em TV-AD, a similaridade semântica é substancialmente maior, sugerindo que as saídas do SONA frequentemente descrevem o mesmo núcleo visual das referências, embora com formas superficiais diferentes. Uma explicação plausível é que os intervalos do TV-AD são curtos e centrados em ações, o que aproxima a referência profissional de uma sentença compacta do tipo “quem fez o quê”. Esse formato é mais próximo da unidade textual produzida pelo compositor do

Tabela 3: Exemplos representativos da comparação entre a audiodescrição profissional e a saída do SONA em *Night of the Living Dead*. Os tempos da referência profissional correspondem às marcações disponíveis no roteiro, enquanto os tempos do SONA correspondem aos cues do WebVTT gerado.

AD profissional	Gerado pelo SONA
1:33:34: “Ben is at the window, his rifle aimed.”	1:33:28–1:33:32: “He holds a firearm and aims through a broken window.”
05:23: “Some distance away, a male figure staggers past a headstone. John dons gloves. He turns back to Barbara.”	05:21–05:26: “He adjusts his gloves, stands with hands on his hips, and looks around.”
No trecho correspondente, a referência mostra a sequência do noticiário.	59:58–1:00:00: “Alex walks toward the door.”
1:25:12: “Repeatedly, Karen plunges the tool into her mother’s neck and torso.”	Nenhum <i>cue</i> correspondente. O <i>cue</i> anterior termina em 1:24:58 e o seguinte começa em 1:25:23.

SONA do que descrições de filmes com maior dependência de contexto narrativo amplo.

5.1.1 Exemplos qualitativos e análise de erros. As métricas agregadas não mostram como as diferenças entre a referência profissional e as saídas do SONA se manifestam em casos individuais. Para complementar essa análise, inspecionamos trechos representativos de *Night of the Living Dead*, incluindo tanto concordâncias quanto diferentes modos de falha. A Tabela 3 apresenta quatro desses casos. Como os roteiros foram originalmente produzidos em inglês, os trechos são mantidos no idioma original para evitar que a tradução introduza diferenças adicionais.

O primeiro caso mostra uma situação de concordância semântica clara. Embora a referência profissional identifique Ben e use “rifle”, enquanto o SONA emprega uma referência pronominal e “firearm”, ambos descrevem a mesma ação central: um homem armado mirando através da janela. Esse caso ilustra por que diferenças de superfície não implicam necessariamente divergência de conteúdo e ajuda a interpretar a diferença observada entre as métricas lexicais e semânticas.

O segundo caso evidencia uma limitação de seleção narrativa. O SONA identifica corretamente que John ajusta as luvas, mas completa a descrição com ações de menor importância, como permanecer com as mãos nos quadris e olhar ao redor. A referência profissional, por outro lado, destaca a figura que cambaleia ao fundo e o retorno de John para Barbara. A divergência, portanto, não decorre apenas de erro factual, mas da escolha de uma unidade de ação menos relevante para a progressão narrativa.

O terceiro caso apresenta um problema de referência a personagens. Durante a sequência do noticiário, o SONA produz “Alex walks toward the door”. O nome “Alex” não aparece no roteiro profissional e a ação não possui correspondência na descrição do mesmo trecho. A partir apenas das saídas textuais, não é possível determinar se o problema se origina

no reconhecimento de personagens ou na inferência multimodal, mas o exemplo evidencia que a nomeação automática pode introduzir referências não sustentadas pelo contexto.

O quarto caso ilustra uma omissão em uma cena crítica. A referência profissional descreve Karen atacando repetidamente Helen em 1:25:12, enquanto o roteiro final do SONA não contém um *cue* correspondente ao evento. Como o roteiro final é resultado tanto da geração por plano quanto da filtragem e alocação pelo Orquestrador, a ausência não permite, isoladamente, atribuir a falha a um único módulo. Ainda assim, ela evidencia que uma medida de similaridade calculada sobre o roteiro completo pode mascarar perdas locais narrativamente significativas.

Outros problemas aparecem mesmo quando uma descrição individual é plausível. Em planos adjacentes, por exemplo, o SONA produz para 00:07:03 “She touches the stone monument with both hands” e, para poucos segundos depois, em 00:07:13, “She grips the stone monument while looking up”. As sentenças descrevem estados muito semelhantes, evidenciando redundância decorrente do processamento local por plano e da ausência de um mecanismo explícito de memória discursiva.

O ajuste temporal também pode introduzir perdas de informação. Em um dos planos, o Compositor originalmente produz “The man looks up at the woman, his serious expression shifting to concern as he holds something in his hands”. Para adequar a narração ao intervalo disponível, o Orquestrador reduz a sentença a “The man holds something while looking at the woman”. O núcleo da ação é preservado e a sentença passa a satisfazer a restrição temporal, mas a mudança de expressão de seriedade para preocupação é removida. Esse exemplo explicita um *trade-off* do mecanismo *fit-to-gap*: a adequação temporal pode ser obtida ao custo de informação visual semanticamente relevante.

Em conjunto, os exemplos mostram que similaridade textual, correção visual, seleção narrativa e adequação temporal constituem dimensões distintas da qualidade de uma audiodescrição. Eles também ajudam a explicar por que as métricas semânticas apresentam resultados mais favoráveis que as métricas lexicais, ao mesmo tempo em que nenhuma delas captura integralmente erros locais de referência, redundância, omissão ou perda de informação causada pelo encurtamento.

5.1.2 Ablação de componentes. A Tabela 2 indica alinhamento com referências profissionais, mas não isola os componentes responsáveis pelos resultados nem mede diretamente o comportamento temporal. Por isso, realizamos uma ablação direcionada sobre dois elementos centrais da proposta: marcação de personagens e ajuste ao intervalo. A ablação textual usa um subconjunto de clipes TV-AD; a ablação temporal usa uma cena de *The Truman Show*. O objetivo não é substituir uma avaliação humana completa, mas verificar se esses componentes contribuem para os dois eixos que motivam o desenho do sistema: descrição semanticamente adequada e narração temporalmente viável.

Tabela 4: Ablação textual sobre cliques TV-AD. BERTScore vem apenas da avaliação automática do texto.

Método	Marcação de Personagens	Ajuste Temporal	BERTScore (F1)
SONA completo	✓	✓	88,29%
SONA s/ marcação	✗	✓	87,92%
SONA s/ ajuste	✓	✗	88,29%
VLM→LLM ingênuo	✗	✗	85,68%

Tabela 5: Ablação temporal em um trecho de *The Truman Show*. Dos planos segmentados até a descrição sintetizada, há duas etapas: a primeira corresponde à geração de uma descrição e a segunda corresponde ao encaixe temporal do áudio sintetizado. A coluna “Adequação” diz respeito às descrições que passaram pelas duas etapas e cujos áudios sintetizados encaixaram no intervalo disponível.

Método	% na etapa 1	% na etapa 2	Adequação	Excesso médio (s)
SONA completo	52,78%	13,89%	100,0%	-
SONA s/ marcação	36,11%	13,89%	100,0%	-
SONA s/ ajuste	52,78%	13,89%	80,0%	1,78
VLM→LLM ingênuo	22,22%	8,33%	0,0%	2,08

A Tabela 4 mostra que o marcação de personagens traz um ganho semântico pequeno, mas consistente: o SONA completo atinge BERTScore (F1) de 88.29%, contra 87.92% sem marcação. O ganho é modesto porque muitos cliques curtos podem ser descritos de forma aceitável sem nomes estáveis, especialmente quando a ação é simples ou envolve poucos personagens. Ainda assim, a diferença sugere que identidade visual explícita ajuda o modelo a manter referências mais coerentes. Como esperado, remover o ajuste temporal não altera esse escore textual, pois a mudança ocorre depois da geração da sentença, na etapa de controle de duração.

Já a Tabela 5 evidencia que o ajuste ao intervalo é o componente temporal crítico. Sem ele, o sistema ainda gera descrições semanticamente equivalentes, mas passa a aceitar narrações que podem ultrapassar a lacuna disponível. Na cena avaliada, sua remoção reduz o sucesso de 100,0% para 80,0% e introduz uma violação média de 1,78s. A linha VLM→LLM ingênua é pior nos dois eixos, com menor similaridade textual e falha completa de ajuste entre as entradas sintetizadas. Isso reforça que a contribuição do SONA não está apenas em consultar um VLM e um LLM, mas em organizar a saída desses modelos dentro das restrições temporais da audiodescrição.

Os denominadores devem ser lidos com cautela. A evidência textual é feita sobre um *dataset* de cliques variados, mas a evidência temporal é menor e específica de uma cena. Além disso, apenas entradas que sobreviveram à alocação

em intervalos e à filtragem de colisões foram sintetizadas. Portanto, a ablação temporal deve ser interpretada como evidência de componente, não como estimativa geral de desempenho temporal em filmes completos. Mesmo assim, ela ilustra o papel prático do mecanismo *fit-to-gap*: sem medir a fala sintetizada, uma descrição curta no texto ainda pode ser longa demais no áudio.

5.2 Avaliação qualitativa

A avaliação automática e a ablação ainda não capturam propriedades centrais de AD, como sincronia percebida, preservação de diálogo e trilha sonora, clareza do roteiro e utilidade prática para edição. Para complementar esses resultados, avaliamos saídas completas do SONA em *Night of the Living Dead* (1968) e *The Truman Show* (1998). O material incluiu o roteiro temporizado e a faixa sintetizada de AD. Um consultor de audiodescrição com 10 anos de experiência assistiu aos dois filmes completos e respondeu a uma avaliação estruturada, registrada de forma agregada para os dois títulos.

A avaliação da faixa mixada foi positiva em propriedades perceptuais básicas. A especialista atribuiu o nível mais favorável à velocidade de leitura e à inteligibilidade, e avaliou o respeito aos intervalos de diálogo e a preservação de efeitos sonoros e música em nível alto, ainda que não máximo. Esses resultados são relevantes porque um sistema de AD não pode ser avaliado apenas pelo texto: uma descrição correta pode comprometer a experiência se for lida rápido demais, se mascarar fala ou se competir com sons diegéticos importantes. No material avaliado, não foi relatado mascaramento relevante da trilha original pela narração gerada.

A sincronia temporal, tanto em relação ao momento dos eventos visuais quanto aos limites de início e fim das narrações, ficou em nível intermediário. Esse resultado é coerente com o desenho atual da *pipeline*. O *fit-to-gap* garante que a fala caiba no intervalo detectado, mas não garante que a sentença descreva o evento visual exatamente no melhor momento narrativo. Essa diferença entre caber no intervalo e estar editorialmente bem posicionado é importante: a primeira propriedade pode ser verificada automaticamente por duração medida; a segunda ainda depende de julgamento contextual e de compreensão narrativa de longo prazo.

O roteiro textual foi avaliado mais favoravelmente como artefato de rascunho do que como produto editorial final. Ele foi considerado claro, conciso, fluente para narração e útil como base de trabalho para um audiodescritor. Controle de redundância e consistência referencial também receberam avaliações favoráveis, o que sugere que a geração por plano não inviabiliza coerência local. Por outro lado, relevância e precisão da informação, seleção narrativa, omissões em cenas críticas, neutralidade descritiva e aderência a convenções estilísticas ficaram em nível intermediário. Esses aspectos exigem decisões editoriais mais finas do que simplesmente reconhecer objetos ou ações.

Em termos práticos, o especialista considerou o material muito útil para aumentar a eficiência de autoria de AD, mas

ainda com esforço moderado de pós-edição. Comentários abertos destacaram a qualidade das descrições e a velocidade de leitura como pontos fortes, e apontaram tempo de processamento como prioridade de melhoria. Assim, os resultados posicionam o SONA como uma ferramenta de apoio e rascunho, não como um sistema autônomo pronto para produção. Esse posicionamento é importante para o escopo do trabalho: a automação proposta reduz parte do custo inicial de autoria, mas não substitui a revisão humana em obras que exigem qualidade editorial profissional.

5.3 Limitações e ameaças à validade

Os resultados devem ser interpretados considerando algumas limitações. Primeiro, os experimentos foram executados sob ambiente computacional restrito, com três quadros por plano e um VLM compatível com uma GPU de uso doméstico. Portanto, o desempenho reportado provavelmente reflete uma configuração conservadora da *pipeline*, não seu limite superior. Modelos maiores, maior densidade de quadros e janelas temporais mais longas podem alterar tanto a qualidade textual quanto a consistência narrativa, embora também aumentem custo e latência.

Segundo, as métricas automáticas comparam a saída gerada com uma única referência profissional, embora a audiodescrição admita múltiplas formulações corretas para o mesmo conteúdo visual. Isso penaliza variações legítimas de estilo, granularidade e seleção narrativa. Além disso, uma pontuação semântica alta não garante *grounding* visual correto: a frase pode ser próxima da referência e ainda errar personagem, objeto, direção de ação ou momento narrativo. Por esse motivo, os resultados textuais devem ser lidos como indícios de alinhamento, não como validação completa de qualidade de AD.

Terceiro, a avaliação temporal da ablação foi conduzida em um trecho específico e com alguns segmentos sintetizados após filtragem por lacunas e colisões. Ela demonstra o efeito do mecanismo de ajuste ao intervalo, mas não deve ser interpretada como estimativa estatística de desempenho temporal em longas-metragens completos. De modo semelhante, a avaliação qualitativa foi realizada por um especialista e registrada de forma agregada para dois filmes. Embora forneça evidência inicial sobre utilidade prática, inteligibilidade e esforço de pós-edição, estudos futuros devem incluir múltiplos avaliadores, análise por cena e usuários cegos ou com baixa visão.

Por fim, a garantia temporal do SONA é condicionada à qualidade dos intervalos detectados pelo VAD. O sistema impede, por construção, que a duração sintetizada exceda a lacuna atribuída; entretanto, se o VAD classificar erroneamente fala residual como silêncio, ainda pode haver sobreposição perceptual. Essa distinção é central: o SONA garante ajuste à lacuna detectada, não perfeição perceptual da segmentação de áudio. A robustez da detecção de fala, a modelagem de sons relevantes não vocais e o suporte a revisão manual dos intervalos permanecem pontos importantes para evolução do sistema.

6 CUIDADOS ÉTICOS

A avaliação qualitativa apresentada neste trabalho foi realizada por meio de questionário eletrônico com um consultor de audiodescrição. Antes de responder ao questionário, o participante teve acesso a um Termo de Consentimento Livre e Esclarecido (TCLE), que informava os objetivos do estudo, o caráter voluntário da participação, a confidencialidade das respostas e a possibilidade de desistência a qualquer momento. A participação somente prosseguiu após manifestação expressa de concordância. O formulário foi conduzido de forma anônima, sem coleta de identificadores diretos, e os resultados são apresentados apenas no contexto da avaliação dos artefatos produzidos pelo SONA. A atividade não foi submetida ao sistema CEP/CONEP, por ter sido conduzida como consulta escrita pontual e anônima para avaliação dos artefatos gerados pelo sistema, conforme a hipótese de pesquisa de opinião pública com participante não identificado prevista na Resolução CNS n.º 510/2016.

7 CONSIDERAÇÕES FINAIS

Este artigo apresentou o SONA, uma *pipeline* híbrida e fim-a-fim para geração de artefatos de audiodescrição diretamente a partir do arquivo audiovisual final. Diferentemente de abordagens que dependem de roteiros, legendas ou treinamento específico, o SONA opera sobre vídeo e áudio renderizados, segmenta o conteúdo em plano, extrai e marca personagens, infere ambiente e eventos visuais e compõe unidades narrativas curtas posteriormente alinhadas a intervalos sem fala. Sua contribuição operacional central é o ajuste explícito ao intervalo durante a síntese, que aceita apenas narrações cuja duração medida respeita a lacuna de silêncio atribuída.

A avaliação mostra um quadro misto, mas coerente. Métricas lexicais subestimam a qualidade de descrições válidas, enquanto métricas semânticas indicam alinhamento mais forte em cenários curtos e centrados em ações. A ablação mostra um ganho semântico modesto com marcação de personagens e evidencia o papel do ajuste temporal na redução de violações de intervalo. A avaliação qualitativa, por sua vez, indica boa inteligibilidade e utilidade prática como rascunho, mas também aponta limitações em sincronia fina, priorização narrativa, omissões, neutralidade e aderência estilística. Em conjunto, os resultados sustentam o SONA como ferramenta de apoio ao fluxo de autoria de AD, mais do que como solução totalmente autônoma.

Trabalhos futuros devem avançar em duas direções. Primeiro, são necessários estudos com múltiplos avaliadores especializados e, idealmente, com usuários cegos ou com baixa visão, para medir factualidade, utilidade e aceitabilidade temporal em condições reais de consumo. Segundo, a *pipeline* pode se beneficiar de melhor *grounding* multimodal, contexto temporal além da tomada isolada, controle mais forte de redundância e priorização de eventos, além de melhorias de robustez em detecção de intervalos e opções de pós-edição. Também é necessário reduzir tempo de processamento para aproximar o sistema de fluxos reais de produção audiovisual.

AGRADECIMENTOS

Os autores agradecem o apoio recebido da FAPERJ, CAPES, FINEP, CNPq e RNP. Ferramentas de Inteligência Artificial Generativa foram utilizadas para revisão linguística. As decisões sobre o conteúdo científico, a metodologia, a análise e interpretação dos resultados, bem como a revisão e aprovação da versão final do texto, são de responsabilidade dos autores.

REFERÊNCIAS

- [1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, 65–72. <https://aclanthology.org/W05-0909>
- [2] G. Bradski. 2000. The OpenCV Library. *Dr. Dobbs's Journal of Software Tools* (2000).
- [3] Virginia Pinto Campos, Tiago Maritan U de Araújo, and Guido L de Souza Filho. 2014. CineAD: Um Sistema de Geração Automática de Roteiros de Audiodescrição. In *Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*. SBC, 1–8.
- [4] Brandon Castellano. 2020. Pyscenedetect. *Last accessed* (2020).
- [5] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*.
- [6] FFmpeg Developers. 2025. ffmpeg tool. <http://ffmpeg.org/>. Version 8.0 “Huffman” [Software].
- [7] Hexgrad. 2025. Kokoro-82M: Open-weight text-to-speech model. <https://huggingface.co/hexgrad/Kokoro-82M>. doi:10.57967/hf/4329 Apache-2.0 License.
- [8] James Lakritz and Andrew Salway. 2006. The semi-automatic generation of audio description from screenplays. *Dept. of Computing Technical Report CS-06-05, University of Surrey* (2006).
- [9] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*. Association for Computational Linguistics, 25–26. <https://aclanthology.org/W04-1013>
- [10] Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.
- [11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 311–318. doi:10.3115/1073083.1073135
- [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] <https://arxiv.org/abs/2103.00020>
- [13] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 3982–3992. doi:10.18653/v1/D19-1410
- [14] James Robert, Marc Webbie, et al. 2018. Pydub. <http://pydub.com/>
- [15] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A Dataset for Movie Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3202–3212. doi:10.1109/CVPR.2015.7298940
- [16] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11987–11997.
- [17] Ollama Team. 2025. Ollama. <https://ollama.com>. Version 0.13.0.
- [18] Silero Team. 2024. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. <https://github.com/snakers4/silero-vad>.
- [19] Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.
- [20] W3C. 2019. *WebVTT: The Web Video Text Tracks Format*. Technical Report W3C Recommendation. World Wide Web Consortium. <https://www.w3.org/TR/webvtt1/>
- [21] Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Eshika Khandelwal, Gül Varol, Weidi Xie, and Andrew Zisserman. 2025. Shot-by-Shot: Film-Grammar-Aware Training-Free Audio Description Generation. *arXiv preprint arXiv:2504.01020* (2025).
- [22] Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. Autoad-zero: A training-free framework for zero-shot audio description. In *Proceedings of the Asian Conference on Computer Vision*. 2265–2281.
- [23] Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. MMAD: Multi-modal movie audio description. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 11415–11428.
- [24] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net. <https://openreview.net/forum?id=SkeHuCVFDr>