

Impacto do tipo de áudio na dispersão dos pontos de fixação em vídeos imersivos: evidências para o desenvolvimento de modelos de saliência visual

Frank Fábio S. da Silva, Saulo G. Martins, Felipe M. Ximenes,
Franciane S. de Azevedo, Ronaldo F. Zampolo e Aline F. G. Sousa

Resumo— Este trabalho investiga o impacto de diferentes condições acústicas (*Ambisonics*, *Mono* e *Mute*) na atenção visual em vídeos 360°. A análise utilizou dados de rastreamento ocular de 60 observadores da base AVS-ODV. Os resultados evidenciam que a presença de áudio reduz significativamente a dispersão dos pontos de fixação, promovendo maior convergência inter-observador, não sendo observadas diferenças estatísticas expressivas entre *Ambisonics* e *Mono*. Conclui-se que, para modelar computacionalmente a saliência audiovisual, o áudio espacial pode não ser estritamente necessário para capturar a convergência global do olhar, viabilizando representações acústicas mais simples.

Palavras-Chave— Atenção Visual, Vídeos 360°, Saliência Audiovisual, Áudio Espacial, Dispersão de Fixações.

Abstract— This work investigates the impact of different acoustic conditions (*Ambisonics*, *Mono*, and *Mute*) on visual attention in 360° videos. The analysis utilized eye-tracking data from 60 observers from the AVS-ODV dataset. The results show that the presence of audio significantly reduces fixation point dispersion, promoting greater inter-observer convergence, with no significant differences observed between *Ambisonics* and *Mono*. Our results indicate that, for the computational modeling of audiovisual saliency, spatial audio may not be strictly necessary to capture the global convergence of fixations, enabling simpler acoustic representations.

Keywords— Visual Attention, 360° Videos, Audio-visual Saliency, Spatial Audio, Fixation Dispersion.

I. INTRODUÇÃO

A crescente adoção de vídeos omnidirecionais (ODV), também denominados *vídeos 360°*, tem transformado a forma como usuários interagem com conteúdos digitais. Diferentemente das mídias tradicionais, em que o consumo é predominantemente passivo, os ODVs permitem exploração visual livre em um ambiente esférico, conferindo ao usuário controle contínuo sobre o campo de visão. Essa característica impõe novos desafios à área de visão computacional, uma vez que compreender e prever a atenção visual torna-se fundamental

para aplicações como *streaming* adaptativo baseado em *viewport* (*viewport-adaptive streaming*) e para a melhoria da Qualidade de Experiência (QoE) em sistemas de Realidade Virtual (RV) [1].

Nesse contexto, modelos de predição de saliência visual desempenham um papel central, pois buscam estimar as regiões de maior interesse visual em uma cena. Em ambientes imersivos, esses modelos são essenciais para otimização de largura de banda, renderização adaptativa e compressão eficiente de conteúdo [2], [3], [4], [5]. Mais recentemente, modelos de saliência audiovisual têm sido propostos [6], [7], [8] com o objetivo de incorporar informações sonoras ao processo de predição, refletindo de forma mais fiel o comportamento humano. No entanto, o desenvolvimento desses modelos ainda levanta uma questão fundamental: qual é, de fato, o papel do áudio na convergência e dispersão da atenção visual?

A percepção audiovisual desempenha um papel central na orientação do comportamento do observador. Devido à limitação intrínseca do campo de visão humano, o sistema auditivo atua como mecanismo complementar na exploração em 360°, capaz de induzir respostas motoras e redirecionar a atenção para fora da visão imediata, ampliando a imersão [9]. Assim, a integração audiovisual torna-se fundamental para compreender a dinâmica atencional em ambientes imersivos.

Modelos computacionais de saliência audiovisual, como o AVS360 [6] e o SVGC-AVA [7], buscam prever de forma eficaz os pontos de fixação mais prováveis utilizando estímulos sonoros. No entanto, essas abordagens priorizam o áudio espacial (*Ambisonics*), sem avaliar se representações mais simples, como *Mono*, produzem efeitos semelhantes. Consequentemente, permanece pouco investigado como a variação no tipo de áudio afeta a dispersão espacial dos pontos de fixação.

Essa lacuna é especialmente relevante para o desenvolvimento de modelos de saliência audiovisual, uma vez que já é estabelecido na literatura que a incorporação de informações sonoras pode melhorar o desempenho desses modelos [8]. O estudo original da base AVS-ODV [10] investigou as mesmas condições acústicas, mas limitou-se a avaliar a cobertura espacial global da cena. A distinção metodológica central reside no objeto de análise, uma vez que a referência se baseou em mapas de saliência agregados. Tais representações estáticas acumulam as fixações de todo o vídeo em uma única imagem e ocultam a evolução temporal, tendo sido utilizadas para a aplicação de métricas de similaridade como o Coeficiente de

Frank Fábio S. da Silva, Saulo G. Martins, Felipe M. Ximenes, Franciane S. de Azevedo e Aline F. G. Sousa; Grupo de pesquisa em Visão Computacional e Aprendizado de Máquina, Faculdade de Engenharia da Computação, Instituto de Geociências e Engenharia, Universidade Federal do Sul e Sudeste do Pará, Marabá-PA, Brasil, e-mails: frankfabio@unifesspa.edu.br, saulo.gomesmartins@unifesspa.edu.br, felipemattei@unifesspa.edu.br, franaezevedo@unifesspa.edu.br, alinefarias@unifesspa.edu.br; Ronaldo F. Zampolo, Laboratório de Processamento de Sinais, Faculdade de Engenharia da Computação e Telecomunicações, Programa de Pós-Graduação em Engenharia Elétrica, Instituto de Tecnologia, Universidade Federal do Pará, Belém-PA, Brasil, email: zampolo@ufpa.br.

Correlação (CC) e a Divergência de *Kullback-Leibler* (D_{KL}). Para superar essa limitação, esta pesquisa inova ao focar na convergência inter-observador dinâmica. Para isso, calcula-se a dispersão a partir das coordenadas exatas de fixação *frame a frame*, mensurando diretamente o quanto os observadores compartilham a mesma região de atenção a cada instante.

Aplicando essa nova abordagem ao conjunto de dados AVS-ODV [1], este trabalho analisa o impacto das condições espacial (*Ambisonics*), não espacial (*Mono*) e de ausência de áudio (*Mute*) sobre a distribuição do olhar. Complementarmente, a similaridade estrutural da atenção visual é avaliada por meio do CC e da D_{KL} , permitindo uma análise integrada dos efeitos do som. O objetivo é fornecer evidências quantitativas sobre como a variação sonora influencia a dispersão das fixações, auxiliando na definição das representações acústicas mais adequadas para o desenvolvimento de modelos preditivos. As principais contribuições deste trabalho são:

- O cálculo da dispersão dos pontos de fixação realizado diretamente em coordenadas esféricas, adotando distância geodésica (*Haversine*) e métrica RMS (*Root Mean Square*);
- A aplicação do Coeficiente de Correlação (CC) e da Divergência de *Kullback-Leibler* (D_{KL}) para quantificar a similaridade entre mapas de saliência obtidos experimentalmente sob diferentes condições de áudio;
- Uma análise sistemática do impacto de diferentes condições de áudio (*Mute*, *Mono* e *Ambisonics*) na dispersão dos pontos de fixação em vídeos 360°, apoiada por testes de significância estatística;
- Evidências empíricas de que a presença de áudio reduz significativamente a dispersão dos pontos de fixação, independentemente das diferentes condições de áudio;
- Uma análise comparativa que demonstra que o áudio ambisônico não apresenta vantagens significativas sobre o áudio mono na convergência da atenção visual;
- Implicações para o desenvolvimento de modelos de saliência audiovisual, indicando que representações de áudio mais simples podem ser suficientes para capturar os principais efeitos do som na organização da atenção visual.

II. METODOLOGIA

A. Base de dados AVS-ODV

Neste trabalho, foi utilizado o conjunto de dados AVS-ODV [1], empregado em estudos de atenção visual em vídeos omnidirecionais. O dataset é composto por 162 vídeos 360° com resolução original de 7680×3840 pixels. Cada vídeo tem duração de aproximadamente 15 segundos, totalizando 450 frames por vídeo (30 FPS).

Para fins de processamento, os frames dos vídeos foram convertidos para o formato equirretangular com resolução de 720×360 pixels. O conjunto também inclui mapas de fixação e mapas de saliência derivados de dados de rastreamento ocular, além dos respectivos arquivos de áudio no formato *.wav*.

Os dados de rastreamento ocular foram coletados a partir de 60 participantes com acuidade visual normal ou corrigida,

utilizando dispositivos de Realidade Virtual. Durante o experimento, os participantes foram divididos em três grupos independentes de 20 indivíduos, cada um associado a uma condição específica de áudio: *Mute*, *Mono* e *Ambisonics*. A ordem de apresentação foi aleatória para reduzir efeitos de ordem e viés de memória.

A Figura 1 apresenta um exemplo de frame extraído de um dos vídeos do conjunto de dados, juntamente com seus respectivos mapas nas três condições.

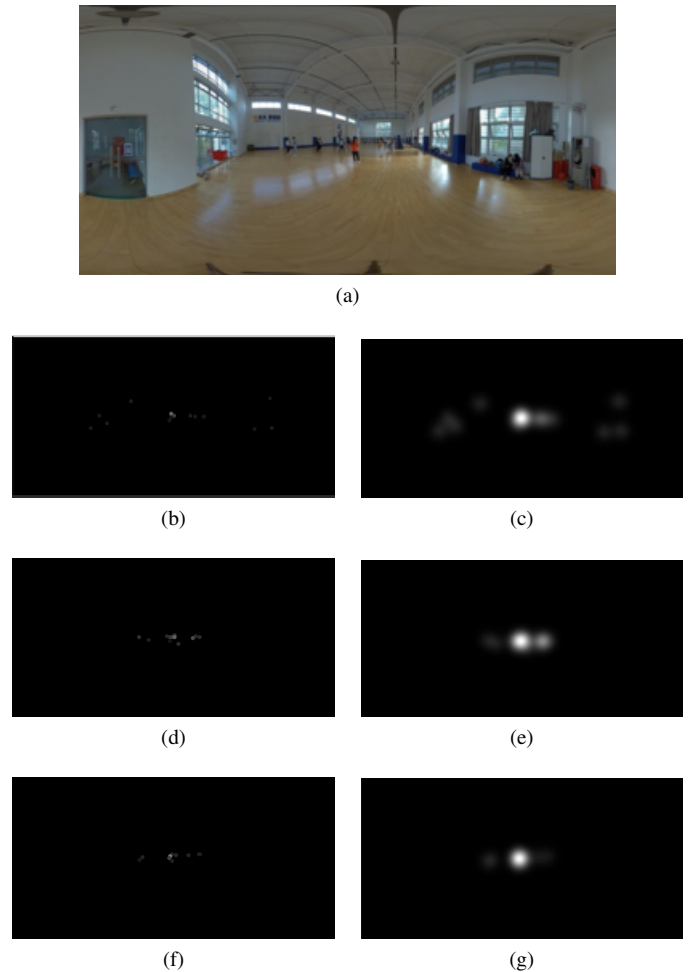


Fig. 1. Exemplo de mapas de saliência visual da base AVS-ODV [10], [1]. Em (a), apresenta-se um frame equirretangular original. As imagens subsequentes ilustram respectivamente os mapas de fixação e os mapas de saliência sob as condições: (b) e (c) *Mute*; (d) e (e) *Mono*; (f) e (g) *Ambisonics*.

Os mapas de fixação foram utilizados para o cálculo da dispersão das fixações, e os mapas de saliência para o cálculo do Coeficiente de Correlação (CC) e da Divergência de *Kullback-Leibler* (D_{KL}).

B. Cálculo da dispersão dos pontos de fixação

Os dados do conjunto AVS-ODV[1] são disponibilizados originalmente em projeção equirretangular (ERP). Para mitigar as distorções geométricas inerentes a esse tipo de projeção, as quais superestimam significativamente as distâncias euclidianas nas regiões polares [11], as posições bidimensionais de fixação foram convertidas para coordenadas angulares esféricas (latitude e longitude).

Após esse ajuste, calculou-se a dispersão dos pontos de fixação frame a frame, aferindo o espalhamento em torno da atenção coletiva. Devido à natureza geométrica do ambiente 360° e à circularidade da longitude, as coordenadas do centroide espacial foram computadas utilizando a média circular para a longitude (λ) e a média aritmética para a latitude (ϕ).

Para mensurar com precisão a distância geodésica entre as coordenadas de cada fixação e o centroide no domínio esférico, adotou-se a fórmula de *Haversine* (d). Essa formulação, que garante estabilidade numérica para ângulos pequenos, é dada pela Equação 1:

$$d_i = 2 \arcsin \left(\sqrt{\sin^2 \left(\frac{\Delta\phi}{2} \right) + \cos \phi_i \cos \bar{\phi} \sin^2 \left(\frac{\Delta\lambda}{2} \right)} \right) \quad (1)$$

Por fim, a dispersão representativa do frame foi consolidada na forma de grandeza quadrática através da métrica RMS. Esse cálculo é formalizado pela Equação 2:

$$D_{RMS} = \sqrt{\frac{1}{N} \sum_{i=1}^N d_i^2}, \quad (2)$$

onde D_{RMS} denota a dispersão dos pontos de fixação por frame expressa em graus, N representa o número total de fixações no frame específico, e d_i denota a distância de Haversine entre a i -ésima fixação e o centroide da atenção do grupo.

C. Cálculo do coeficiente de correlação

Para complementar a análise da dispersão dos pontos de fixação e avaliar a similaridade estrutural entre os mapas de saliência obtidos nas diferentes condições de áudio (*Mute*, *Mono* e *Ambisonics*), foi calculado o Coeficiente de Correlação Linear de Pearson (CC).

Essa métrica permite quantificar o grau de associação linear entre duas distribuições, sendo amplamente utilizada para comparar mapas de saliência visual.

Valores de CC próximos de 1 indicam alta similaridade estrutural entre os padrões analisados, enquanto valores próximos de 0 indicam ausência de correlação. O coeficiente foi calculado conforme definido na Equação 3:

$$CC = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}}, \quad (3)$$

onde X e Y representam os mapas de saliência comparados, X_i e Y_i denotam o valor de saliência do índice individual de cada elemento (pixel) dos respectivos mapas, $\bar{X}$ e $\bar{Y}$ são as suas médias espaciais, e N corresponde ao número total de elementos (pixels).

D. Divergência de Kullback-Leibler

Enquanto o CC avalia a correlação linear entre os mapas, a Divergência de Kullback-Leibler (D_{KL}) quantifica o afastamento estatístico e as assimetrias na distribuição da atenção visual sob as diferentes condições experimentais. Por ser uma

métrica não simétrica, o cálculo de X para Y difere de Y para X . Nessa métrica, valores próximos a 0 indicam distribuições altamente similares, crescendo proporcionalmente ao aumento da divergência espacial. A D_{KL} do mapa de saliência Y em relação ao mapa de referência X é definida pela Equação 4:

$$D_{KL}(X||Y) = \sum_{i=1}^N X_i \log \left(\frac{X_i}{Y_i + \epsilon} \right), \quad (4)$$

onde X_i e Y_i denotam o valor de saliência do índice individual de cada elemento (pixel) dos respectivos mapas, X e Y , e N representa o número total de elementos. Os valores de X_i e Y_i foram normalizados para representarem distribuições de probabilidade de soma unitária. Adicionalmente, o fator estabilizador $\epsilon = 10^{-12}$ foi incorporado ao denominador para prevenir instabilidades numéricas provocadas por eventuais valores nulos.

III. RESULTADOS E DISCUSSÕES

A análise da dispersão média revelou um padrão consistente em relação ao papel funcional do áudio na orientação da atenção. A Figura 2 ilustra a evolução temporal dessa dinâmica ao longo dos vídeos.

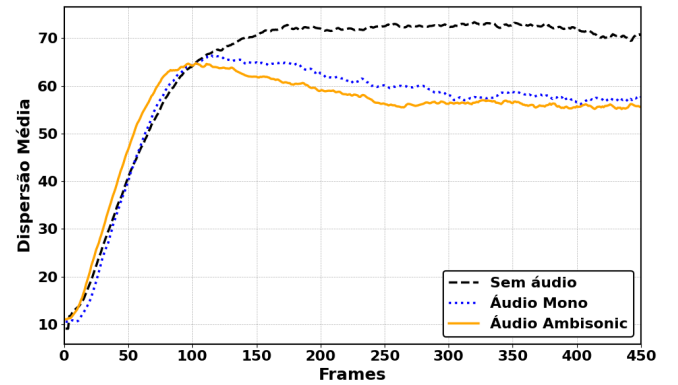


Fig. 2. Evolução temporal da dispersão média dos pontos de fixação.

Devido ao protocolo experimental empregado na coleta da base de dados, todos os observadores iniciam a visualização rigidamente alinhados ao centro do campo visual. Esse fator introduz, nos instantes iniciais, uma elevada similaridade induzida artificialmente. A observação da evolução temporal (Figura 2) permitiu identificar a formação de um platô natural de estabilização da atenção. Com base nessa observação, analisamos os dados a partir do frame 101 dos vídeos. Essa estratégia mitigou o viés de centralização, isolando padrões de atenção visual que refletem puramente a exploração espontânea do ambiente imersivo.

Diante disso, os experimentos foram sistematizados em dois cenários analíticos complementares:

- (i) utilizando a extensão temporal completa (450 frames); e
- (ii) restringindo a análise ao subconjunto validado pelo critério de corte restritivo. A Tabela I consolida esses achados.

TABELA I

 DISPERSÃO MÉDIA GLOBAL DOS PONTOS DE FIXAÇÃO (RMS EM GRAUS)
 NOS CENÁRIOS COMPLETO E COM CORTE.

Condição	Sem corte	Com corte
<i>Mute</i>	64,10	67,69
<i>Mono</i>	55,40	58,51
<i>Ambisonics</i>	54,50	57,27

Os dados atestam que a ausência de estímulos sonoros (*Mute*) resultou na maior dispersão média das fixações. A aplicação do corte restritivo elevou ligeiramente as médias globais em todas as condições, um comportamento esperado, visto que os instantes iniciais de convergência induzida pelo protocolo experimental foram descartados. Mesmo após esse ajuste, as condições dotadas de áudio (*Mono* e *Ambisonics*) mantiveram valores inferiores à condição *Mute*, confirmando o papel do som como um mecanismo de ancoragem atencional.

A Figura 3 aprofunda essa avaliação sob uma perspectiva de densidade.

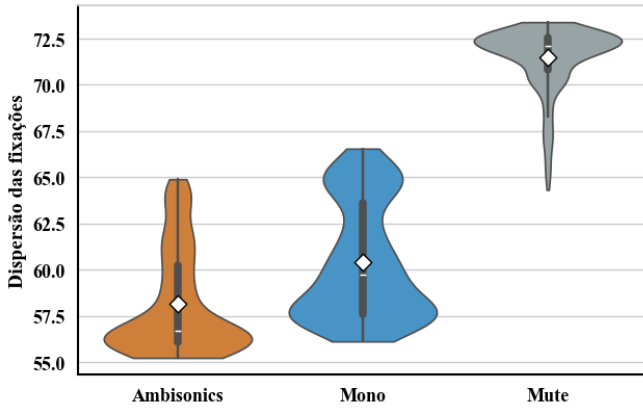


Fig. 3. Distribuição da dispersão das fixações após o corte dos 100 frames iniciais. A remoção do viés de centralização inicial do experimento acentua a visualização da dispersão, especialmente na condição *Mute*.

O gráfico de violino evidencia que a distribuição da condição *Mute* concentra a maior parte do seu volume em valores elevados de dispersão das fixações (acima de 70 graus), indicando uma divergência espacial constante entre os observadores. Em oposição, a introdução sonora desloca o volume principal da distribuição para faixas significativamente mais baixas, concentrando a densidade na base dos gráficos (*Mono* e *Ambisonics*), o que confirma visualmente o efeito do som na redução da dispersão.

A avaliação da significância estatística sobre o cenário com corte confirmou as diferenças observadas. Como o teste de Shapiro-Wilk rejeitou a premissa de normalidade ($p < 0,05$) em um desenho de três grupos independentes, aplicou-se o teste não-paramétrico de Kruskal-Wallis. Os resultados apontaram um efeito principal significativo da condição acústica sobre a dispersão das fixações ($H = 75,58$; $p < 0,001$), demonstrando matematicamente que a presença ou ausência de som afeta de forma direta o espalhamento do olhar.

O detalhamento dessas diferenças entre as condições acústicas foi conduzido pelo teste *post-hoc* de Mann-Whitney,

cujos resultados encontram-se sumarizados na Tabela II. Nessa etapa, a correção de Holm foi empregada para resguardar os valores de p contra falsos positivos decorrentes de múltiplas comparações, enquanto o cálculo do g de Hedges dimensionou a magnitude prática do impacto de cada alteração.

TABELA II

 RESULTADOS DO TESTE POST-HOC (MANN-WHITNEY) ENTRE AS
 CONDIÇÕES DE ÁUDIO

Par comparado	Valor- p corrigido	Tamanho do efeito (g)
<i>Mute</i> $\times$ <i>Mono</i>	$< 0,001$	0,84
<i>Mute</i> $\times$ <i>Ambi</i>	$< 0,001$	0,99
<i>Mono</i> $\times$ <i>Ambi</i>	0,163	0,16

Os resultados dos testes indicam ausência de diferença estatística significativa na dispersão das fixações entre as condições *Ambisonics* e *Mono* ($p = 0,163$; g negligenciável). Em contrapartida, a comparação das condições sonoras com o cenário sem áudio (*Mute*) revela uma redução expressiva da dispersão ($p < 0,001$). A magnitude dessa mudança é sustentada pelos fortes tamanhos de efeito (com g igual a 0,99 e 0,84, em valores absolutos), confirmando o impacto de ancoragem do som na convergência do olhar.

Para complementar a análise e estabelecer uma linha de base comparativa, a similaridade estrutural e a divergência espacial dos mapas de saliência foram quantificadas pelas métricas CC e D_{KL} , respectivamente. As Tabelas III e IV apresentam os resultados obtidos neste estudo em contraste com os valores reportados pelos autores originais da base de dados [10]. Esse paralelo tem o propósito de validar as tendências de atenção visual observadas com a nossa metodologia, alinhadas aos padrões já documentados na literatura.

TABELA III

 COEFICIENTES DE CORRELAÇÃO (CC) NO CENÁRIO COMPLETO, NO
 CENÁRIO COM CORTE (APÓS O FRAME 100) E OS VALORES DA
 REFERÊNCIA [10].

Comparação	Ref. [10]	Sem corte	Com corte
<i>Mute</i> $\times$ <i>Mono</i>	0,777	0,73	0,71
<i>Mute</i> $\times$ <i>Ambi</i>	0,751	0,70	0,69
<i>Mono</i> $\times$ <i>Ambi</i>	0,832	0,78	0,78

TABELA IV

 DIVERGÊNCIA DE KULLBACK-LEIBLER (D_{KL}) NO CENÁRIO COMPLETO,
 NO CENÁRIO COM CORTE (APÓS O FRAME 100) E OS VALORES DA
 REFERÊNCIA [10].

Comparação	Ref. [10]	Sem corte	Com corte
<i>Mute</i> $\times$ <i>Mono</i>	2,856	3,51	3,91
<i>Mute</i> $\times$ <i>Ambi</i>	3,116	3,74	4,16
<i>Mono</i> $\times$ <i>Ambi</i>	2,476	2,63	2,87

Após o corte, observa-se um aumento consistente nos valores de D_{KL} , como esperado, evidenciando de forma mais clara as diferenças reais no comportamento exploratório dos observadores. Ainda assim, o padrão geral se mantém alinhado

à ordenação reportada em [10]. A divergência da condição *Mono* em relação à *Ambi* (2,87) continua significativamente menor do que as divergências da condição *Mute* em relação à *Mono* (3,91) e à *Ambi* (4,16).

Cabe ressaltar que, embora este trabalho tenha avaliado as mesmas condições acústicas e adotado duas das métricas utilizadas em [10] (CC e D_{KL}), os coeficientes de correlação obtidos sob a nossa abordagem metodológica foram ligeiramente inferiores (e os valores de D_{KL} superiores) aos reportados pelos autores originais. Essa diferença observada nos resultados pode ser justificada pela possibilidade de que os experimentos de [10] tenham sido realizados com uma versão diferente da base AVS-ODV [1]. Embora ambos os trabalhos mencionem 162 vídeos, não há confirmação de que a versão utilizada seja exatamente a mesma.

Em conjunto com as análises de dispersão e distribuição, esses resultados indicam que o áudio não altera substancialmente para onde os observadores olham, mas atua principalmente na organização da atenção, reduzindo sua variabilidade e promovendo maior convergência entre as fixações dos indivíduos.

IV. CONCLUSÃO

Este trabalho investigou o impacto de diferentes condições de áudio na dispersão dos pontos de fixação em vídeos 360°, utilizando o conjunto AVS-ODV e aplicando o cálculo de dispersão diretamente em coordenadas esféricas. Apoiados por análises inferenciais (Kruskal-Wallis e Mann-Whitney), os resultados evidenciam que a presença de estímulos auditivos está associada a uma redução significativa da dispersão, promovendo maior convergência da atenção visual em ambientes imersivos. A relevância dessas mudanças, confirmada pelo cálculo de tamanho de efeito (g de Hedges), é corroborada pela análise das distribuições por meio dos gráficos de violino, que mostram uma redução da variabilidade e maior concentração dos valores nas condições com áudio em comparação à condição *Mute*.

Por outro lado, a ausência de diferenças expressivas entre as condições *Mono* e *Ambisonics*, aliada à forte sobreposição de suas distribuições, indica que a contribuição do áudio para a organização da atenção visual é predominantemente guiada por seu conteúdo semântico, e não por sua direcionalidade espacial. Esse resultado é reforçado pelas métricas de comparação (Tabelas III e IV), que evidenciam uma maior correlação e uma menor divergência (D_{KL}) entre as condições *Mono* e *Ambisonics* quando comparadas às demais combinações. Esse comportamento sugere que a estrutura espacial global da atenção visual é amplamente preservada, independentemente do grau de espacialização do áudio.

Como trabalhos futuros, investigações adicionais podem explorar métricas focadas na área de cobertura e na complexidade espacial da atenção, tais como a *Normalized Scanpath Saliency* (NSS), a *Bivariate Contour Ellipse Area* (BCEA), a Entropia Espacial e a *Convex Hull Area*. A aplicação dessas métricas a diferentes categorias de conteúdo e complexidades sonoras ampliará a compreensão da interação audiovisual imersiva.

Do ponto de vista aplicado, esses achados apontam que modelos de predição de saliência audiovisual podem se beneficiar

do uso de representações de áudio mais simples, mantendo a capacidade de capturar os principais efeitos do som na atenção coletiva, ao mesmo tempo em que reduzem custos computacionais.

Em síntese, os resultados indicam que, para a modelagem da atenção visual em vídeos imersivos, a presença do som atua como principal fator de organização e convergência da atenção, enquanto sua espacialização, embora perceptualmente relevante, não se mostra essencial para explicar a convergência coletiva do olhar.

AGRADECIMENTOS

Este estudo tem o suporte financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA). Agradecimentos ao grupo de pesquisa em Visão Computacional e Aprendizado de Máquina (VICAM) da Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA) e ao Laboratório de Processamento de Sinais (LaPS) da Universidade Federal do Pará (UFPA) pelo suporte técnico.

REFERÊNCIAS

- [1] Y. Zhu, H. Duan, K. Zhang, Y. Zhu, X. Zhu, L. Teng, X. Min, and G. Zhai, "How does audio influence visual attention in omnidirectional videos? database and model," *IEEE Transactions on Image Processing*, 2025.
- [2] M. Qiao, M. Xu, Z. Wang, and A. Borji, "Viewport-dependent saliency prediction in 360° video," *IEEE Transactions on Multimedia*, vol. 23, pp. 748–760, 2021.
- [3] C. H. Vo, J.-C. Chiang, D. H. Le, T. T. Nguyen, and T. V. Pham, "Saliency prediction for 360-degree video," in *2020 5th International Conference on Green Technology and Sustainable Development (GTSD)*, pp. 442–448, 2020.
- [4] M. Coklek, N. Imamoglu, C. Ozcinar, E. Erdem, and A. Erdem, "Spherical vision transformer for 360° video saliency prediction," in *34th British Machine Vision Conference (BMVC 2023)*, 2023.
- [5] F. M. Ximenes, J. V. R. Barros, R. F. Zampolo, and A. F. G. Sousa, "Modelagem de atenção audiovisual em vídeos imersivos: impactos do áudio na dispersão ocular," in *Anais do XLIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBt)*, (Natal, Brasil), October 2025.
- [6] F.-Y. Chao, C. Ozcinar, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Towards audio-visual saliency prediction for omnidirectional video with spatial audio," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 355–358, IEEE, 2020.
- [7] Q. Yang, Y. Li, C. Li, H. Wang, S. Yan, L. Wei, W. Dai, J. Zou, H. Xiong, and P. Frossard, "SVGC-AVA: 360-degree video saliency prediction with spherical vector-based graph convolution and audio-visual attention," *IEEE Transactions on Multimedia*, vol. 26, pp. 3061–3076, 2023.
- [8] Z. Wan, H. Qin, Z. Li, X. Fan, W. Zuo, and D. Zhao, "CASP: Consistency-aware audio-induced saliency prediction model for omnidirectional video," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12605–12614, 2025.
- [9] F.-Y. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Audio-visual perception of omnidirectional video for virtual reality applications," in *ICME Workshops*, pp. 1–6, 2020.
- [10] Y. Zhu, X. Zhu, H. Duan, J. Li, K. Zhang, Y. Zhu, L. Chen, X. Min, and G. Zhai, "Audio-visual saliency for omnidirectional videos," in *International Conference on Image and Graphics*, pp. 365–378, Springer, 2023.
- [11] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun, "Cube padding for weakly-supervised saliency prediction in 360° videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1420–1429, 2018.