3.5.3. Efectividad de etiquetas de advertencia: evidencias empíricas

La investigación sobre la efectividad de diferentes formatos de etiquetado ha producido hallazgos matizados:

- **Etiquetas explícitas vs. sutiles**: Las advertencias prominentes son más efectivas para captar la atención, pero pueden generar reactancia o desconfianza generalizada, mientras que los indicadores sutiles preservan la experiencia del usuario, aunque pueden pasar desapercibidos.
- **Momentos de exposición**: La efectividad varía según cuándo se presenta la advertencia (antes, durante o después del consumo del contenido).
- **Fatiga de advertencias**: La exposición repetida a alertas puede llevar a la habituación y a la disminución de la efectividad con el tiempo, particularmente en contextos de alto volumen de contenido.

Los estudios citados por Chesney y Citron [5] sugieren que las etiquetas más efectivas combinan información clara sobre el origen sintético con indicaciones contextuales sobre el propósito de la generación.

3.5.4. Impacto en confianza institucional y salud democrática

La literatura especializada [5,6] identifica múltiples mecanismos a través de los cuales los deepfakes afectan la salud del ecosistema informativo:

- **Erosión de confianza generalizada**: La mera posibilidad de existencia de contenido sintético convincente puede socavar la confianza en material auténtico (efecto "liar's dividend").
- **Polarización y fragmentación**: Los deepfakes pueden exacerbar divisiones sociales al proporcionar "*evidencia*" fabricada que confirma narrativas polarizantes.
- **Alteración de procesos deliberativos**: La intoxicación del espacio informativo con contenido sintético dificulta el debate público basado en hechos compartidos.
- **Costos de verificación**: La necesidad de implementar sistemas de detección y verificación impone cargas significativas a las organizaciones mediáticas, plataformas y usuarios individuales.

Estos impactos sistémicos subrayan la urgencia de desarrollar no solo contramedidas técnicas efectivas, sino también marcos normativos apropiados y programas integrales de alfabetización digital.

4 Metodología del trabajo

4.1 Enfoque general

La investigación adopta un diseño metodológico mixto secuencial explicativo que combina métodos cuantitativos y cualitativos en tres fases interrelacionadas, siguiendo las mejo-

res prácticas establecidas en la literatura sobre evaluación de sistemas interactivos [4] y estudios de percepción mediática [6].

La fundamentación teórico-metodológica se basa en el paradigma pragmático, priorizando la utilidad práctica de los hallazgos y su aplicabilidad en contextos reales de consumo de contenido digital. La Tabla 1 presenta el framework metodológico integral:

Cuadro 1: Framework metodológico integral de la investigación

Fase	Paradigma	Objetivos específicos	Técnicas principales
Fase 1: Desarrollo tecnológico	Investigación-acción	▪ Generar corpus controlado ▪ Implementar etiquetado multimodal ▪ Integrar C2PA ▪ Validar técnicamente	▪ Ingeniería de software ▪ Análisis forense digital ▪ Pruebas de validación ▪ Documentación técnica
Fase 2: Experimentación	Empírico-analítico	▪ Medir percepción humana ▪ Evaluar efectividad etiquetas ▪ Identificar factores moderadores ▪ Analizar procesos decisorios	▪ Experimentos controlados ▪ Medidas psicofísicas ▪ Análisis estadístico ▪ Modelado multinivel
Fase 3: Análisis integrado	Interpretativo	▪ Integrar hallazgos ▪ Desarrollar modelo explicativo ▪ Elaborar recomendaciones ▪ Validar externamente	▪ Triangulación metodológica ▪ Análisis cualitativo ▪ Validación por expertos ▪ Meta-análisis

4.1.1. Justificación de la selección de Runway ML

La elección de Runway ML como herramienta central se fundamenta en criterios técnicos, metodológicos y pragmáticos exhaustivos:

Criterios técnicos:

- **Arquitectura multimodal**: Soporte para text-to-video, image-to-video, y style transfer
- **APIs documentadas**: Acceso programático para automatización y registro de parámetros
- **Reproducibilidad**: Capacidad de replicar generaciones mediante seed control
- **Escalabilidad**: Infraestructura cloud que permite generación masiva controlada

Criterios metodológicos:

- **Representatividad**: Posición como herramienta líder en el ecosistema actual [8]
- **Accesibilidad**: Interfaz que facilita la participación de usuarios no expertos
- **Trazabilidad**: Capacidad de registrar completamente el proceso generativo
- **Compatibilidad**: Integración con estándares de procedencia emergentes

Criterios pragmáticos:

- **Adopción educativa**: Uso documentado en contextos formativos [8]
- **Aplicación profesional**: Implementación en producción de contenido [10]
- **Comunidad activa**: Soporte y documentación extensiva disponible
- **Actualizaciones frecuentes**: Mantenimiento de relevancia tecnológica

4.1.2. Hipótesis y variables de investigación

El estudio plantea las siguientes hipótesis derivadas del marco teórico:

Hipótesis principal:

- **H0**: Los vídeos generados por IA con Runway ML son perceptualmente indistinguibles de los vídeos reales para los observadores humanos en condiciones de visualización cotidianas, superando el umbral de credibilidad crítica.

Hipótesis específicas:

- **H1**: La presencia de etiquetas de procedencia C2PA verificable mejora significativamente la precisión en la detección de contenido sintético generado con Runway ML
- **H2**: Existe un efecto de interacción significativo entre la calidad técnica del contenido y la efectividad de los diferentes tipos de etiquetado
- **H3**: La alfabetización digital media la relación entre la exposición a etiquetas y la confianza subjetiva en los juicios de autenticidad
- **H4**: Las medidas implícitas (como el tiempo de respuesta) revelarán indicios de detección subconsciente de contenido sintético incluso cuando la decisión consciente del participante sea incorrecta.

Matriz de variables operacionalizadas:

Cuadro 2: Operacionalización de variables de investigación

Variable	Tipo	Definición operacional	Instrumento de medida
Precisión detección	Dependiente	Proporción de clasificaciones correctas en tarea forced-choice	Matriz confusión, d', AUC (ROC)
Credibilidad percibida	Dependiente	Puntuación agregada escala Likert 1-7 (7 ítems)	Escala McCroskey adaptada, $\alpha = 0.89$
Confianza subjetiva	Dependiente	Porcentaje auto-reportado 0-100 % por juicio	Escala visual analógica, calibración
Tipo etiquetado	Independiente	Manipulación experimental (3 niveles)	Asignación aleatoria estratificada
Calidad técnica	Independiente	Nivel objetivo (alta/ media/baja)	Métricas PSNR, SSIM, VMAF
Alfabetización digital	Moderadora	Puntuación escala compuesta 1-5	DLI-8 adaptado, $\alpha = 0.85$
Tiempo respuesta	Mediadora	Latencia en milisegundos por juicio	Timestamp high-resolution

4.2 Generación del banco de pruebas

4.2.1. Diseño factorial del corpus de estímulos

El banco de pruebas se construye mediante un diseño factorial completo $2 \times 3 \times 3 \times 3$ que permite el análisis de efectos principales e interacciones de hasta tercer orden. La estructura factorial se detalla en la Tabla 3:

Cuadro 3: Esquema del diseño factorial completo del corpus de estímulos

Factor	Niveles	Operacionalización	Control experimental
Autenticidad	2 niveles	▪ Real: Captura directa ▪ Sintético: Runway ML	Verificación forense, hash criptográfico
Calidad técnica	3 niveles	▪ Alta: 1080p, >8Mbps ▪ Media: 720p, 4-8Mbps ▪ Baja: 480p, <4Mbps	Re-encoding controlado, métricas objetivas
Temática contenido	3 niveles	▪ Informativa ▪ Social ▪ Educativa	Scripts estandarizados, validación inter-jueces
Duración	3 niveles	▪ Corto: 5-15s ▪ Medio: 30-60s ▪ Largo: 90-120s	Edición precisa, normalización temporal

El diseño factorial completo genera $2 \times 3 \times 3 \times 3 = 54$ celdas experimentales. Para asegurar poder estadístico adecuado y robustez, cada celda se replica 3 veces, resultando en un corpus total de 162 estímulos. La asignación de réplicas sigue un diseño balanceado que controla para variabilidad intra-celda.

4.2.2. Proceso técnico de generación con Runway ML

La generación de contenido sintético sigue un protocolo estandarizado que garantiza consistencia, reproducibilidad y documentación completa. La Tabla 4 detalla el proceso:

Cuadro 4: Protocolo técnico de generación de estímulos con Runway ML

Etapa	Actividades	Parámetros controlados	Documentación
1. Preparación	▪ Diseño prompts ▪ Configuración proyecto ▪ Calibración inicial	▪ Modelo: Gen-2 ▪ Resolución: 1024×576 ▪ Aspect ratio: 16:9	Log proyecto, versionado git
2. Generación	▪ Ejecución batches ▪ Control seeds ▪ Monitorización proceso	▪ CFG: 7.5, 12.5, 15.0 ▪ Steps: 25, 50, 100 ▪ Temperature: 0.7-1.2	Logs ejecución, screenshots
3. Selección	▪ Evaluación calidad ▪ Detección artifacts ▪ Balanceo corpus	▪ PSNR >28 dB ▪ SSIM >0.85 ▪ VMAF >85	Matriz selección, criterios aplicados
4. Post-proceso	▪ Normalización ▪ Estandarización ▪ Metadatos	▪ Color: Rec.709 ▪ Audio: -23 LUFS ▪ FPS: 30 constante	Scripts procesamiento, logs

Desarrollo de prompts estructurados:

Los prompts se desarrollan siguiendo una taxonomía jerárquica que asegura cobertura sistemática del espacio semántico:

Cuadro 5: Taxonomía de prompts para generación de contenido

Categoría temática	Subcategorías	Estructura prompt	Ejemplos
Informativa	Noticias, reportajes, declaraciones	[Sujeto] + [acción] + [contexto] + [estilo visual]	"Periodista dando noticias en estudio profesional"
Social	Conversaciones, entrevistas, interacciones	[Participantes] + [situación] + [relación] + [ambiente]	"Dos amigos conversando en café urbano animado"
Educativa	Tutoriales, explicaciones, demostraciones	[Instructor] + [contenido] + [audiencia] + [formato]	"Profesor explicando concepto científico a estudiantes"

Cada prompt se construye mediante combinación sistemática de componentes siguiendo esquemas gramaticales predefinidos. Se generan 5 variantes por combinación temática, con validación inter-jueces para asegurar adecuación y ausencia de sesgos.

Configuración técnica detallada:

Los parámetros técnicos de generación se seleccionan basándose en experimentación preliminar y mejores prácticas documentadas:

Cuadro 6: Configuración técnica detallada de parámetros Runway ML

Parámetro	**Valores**	**Efecto técnico**	**Justificación metodológica**
CFG Scale	7.5, 12.5, 15.0	Control adherencia prompt vs. creatividad	Cobertura espectro conservador-creativo
Number of Steps	25, 50, 100	Iteraciones denoising, detalle vs. tiempo	Balance calidad-computación, artifacts
Seed Control	Fijo por variante	Reproducibilidad generaciones	Control experimental, replicabilidad
Interpolación	Sí/No	Suavizado movimiento, fluidez	Evaluación coherencia temporal
Style Presets	Realista, Cinematic, Artistic	Apariencia visual, estilización	Diversidad perceptual, generalización

4.2.3. Validación técnica exhaustiva del banco de pruebas

La validación del corpus sigue un protocolo multicapa que incluye evaluación automática, análisis por expertos, y pruebas psicofísicas:

Validación automática mediante métricas objetivas:

Cuadro 7: Protocolo de validación automática del banco de pruebas

Dimensión	**Métricas**	**Umbrales aceptación**	**Herramientas**
Calidad visual	PSNR, SSIM, VMAF	PSNR >28 dB, SSIM >0.85, VMAF >80	FFmpeg, VMAF Netflix, MATLAB
Artifacts GAN	FID, KID, Artifact Ratio	FID <50, Artifact Ratio <0.05	Clean-FID, TIMM, DetectGAN
Consistencia temporal	tLPIPS, Warping Error	tLPIPS <0.15, Warping Error <2.5	LPIPS, FlowNet2, RAFT
Análisis forense	Error Level Analysis, Noise Analysis	Consistencia patrones ruido	Forensically, Ghiro, Amped Five

Validación por expertos:

La evaluación por expertos incluye 5 especialistas con experiencia en visión por computadora y análisis forense digital. El protocolo de validación sigue:

Cuadro 8: Protocolo de validación por expertos del banco de pruebas

Criterio	**Escala evaluación**	**Procedimiento**	**Criterios aceptación**
Realismo perceptual	Likert 1-7 (muy artificial - muy realista)	Visualización aleatoria, doble ciego	Media >4.5, σ <1.2
Calidad técnica	Likert 1-7 (muy baja - muy alta)	Evaluación frame-by-frame	Media >5.0, σ <1.0
Artifacts visibles	Checklist 20 ítems	Inspección sistemática por regiones	<3 artifacts críticos por vídeo
Coherencia semántica	Escala 1-5 (incoherente - muy coherente)	Evaluación contenido vs. prompt	Media >4.0, acuerdo >80 %

Pruebas psicofísicas preliminares:

Se realiza un estudio piloto con 20 participantes para validar la discriminabilidad perceptual del corpus:

- **Tarea**: Discriminación forced-choice real vs. sintético
- **Métricas**: d', criterio de respuesta, curvas ROC
- **Criterio**: d' entre 0.8 y 2.5 (discriminabilidad moderada)
- **Ajustes**: Reemplazo de estímulos fuera de rango objetivo

4.3 Desarrollo tecnológico

4.3.1. Arquitectura del sistema de etiquetado multimodal

El sistema de etiquetado implementa una arquitectura modular que integra múltiples capas de información de procedencia. La Tabla 9 describe la implementación:

Cuadro 9: Arquitectura técnica del sistema de etiquetado multimodal

Capa	**Tecnología**	**Implementación**	**Parámetros configurables**
Visible overlay	SVG + CSS animations	Superposición canvas HTML5, composición alpha	Posición, tamaño, opacidad, timing, easing
Marca agua invisible	DCT domain embedding	Modulación coeficientes media frecuencia (canal Y)	Strength (0.1-0.4), robustness, capacity
Metadatos XMP	XML schema extension	Dublin Core + custom fields, sidecar files	Schema validation, namespace management
Código QR dinámico	QR Code + deep linking	URL con session tokens, parameters encoding	Error correction, size, format (PNG/SVG)
Procedencia C2PA	JSON-LD manifests	Claims structure, cryptographic signatures	Hash algorithms, timestamp authorities

Implementación de marcas de agua invisibles:

El algoritmo de marcas de agua implementa técnicas avanzadas de procesamiento de señal:

Cuadro 10: Especificaciones técnicas del sistema de marcas de agua invisibles

Parámetro	**Configuración**	**Algoritmo**	**Robustez verificada**
Domain	DCT (8×8 blocks)	Modulation mid-band coefficients	JPEG compression up to $Q = 70$
Strength	$\alpha = 0.25$	Adaptive based on local complexity	PSNR >40 dB, SSIM >0.98
Capacity	64 bits payload	Error correction (BCH code)	BER $<10^{-4}$ after attacks
Synchronization	Template-based	Auto-correlation peak detection	Rotation ±5°, scaling ±10 %
Detection	Normalized correlation	Neyman-Pearson hypothesis testing	P_{fa} <0.01, P_d >0.95

La implementación incluye:

- Pre-procesamiento: Normalización espacial y espectral
- Inserción: Modulación en dominio de frecuencia con máscara perceptual
- Extracción: Correlación normalizada con umbral adaptativo
- Validación: Pruebas exhaustivas de robustez y imperceptibilidad

4.3.2. Integración avanzada del estándar C2PA

La implementación de C2PA sigue rigurosamente la especificación 1.3 [11] con extensiones para contenido generativo:

Cuadro 11: Implementación detallada del estándar C2PA para procedencia digital

Componente C2PA	**Formato técnico**	**Contenido específico**	**Validación**
Asset Reference	SHA-256 hash	Video file, metadata, thumbnails	Hash chain verification
Claim: Generation	JSON-LD schema	▪ tool: Runway ML Gen-2" ▪ version: "v2.1.0" ▪ timestamp: ISO 8601	Schema.org validation
Claim: Parameters	Structured JSON	▪ prompt: text ▪ cfg_scale: 12.5 ▪ steps: 50 ▪ seed: 123456	Type checking, range validation
Evidence: Logs	GZIP compressed	API responses, generation logs	Temporal consistency checks
Signature	RSA-2048 + PSS	Digital signature over manifest	Certificate chain validation
Timestamp	RFC 3161	Trusted timestamp authority	TSA certificate validation

Arquitectura de verificación C2PA:

El sistema de verificación implementa una pipeline completa:

Cuadro 12: Pipeline de verificación de procedencia C2PA

Etapa	Proceso	Herramientas	Criterios aceptación
1. Extracción	Parse manifest C2PA	c2pa-rs library, custom parser	Structural integrity, schema compliance
2. Validación	Verify signatures, hashes	OpenSSL, cryptographic libraries	Signature valid, hash chain intact
3. Verificación	Check timestamps, certificates	TSA client, certificate validation	Timestamp valid, certs not revoked
4. Interpretación	Render human-readable	React components, i18n	Clear presentation, accessibility
5. Auditoría	Log verification process	Structured logging, analytics	Complete audit trail

Desarrollo de herramientas complementarias:

- **C2PA Validator CLI**: Herramienta línea comandos para validación batch
- **Browser Extension**: Extensión Chrome/Firefox para verificación en tiempo real
- **Mobile SDK**: Librería iOS/Android para aplicaciones móviles
- **API REST**: Servicio web para integración con otras plataformas
- **Dashboard Analytics**: Panel monitorización uso y efectividad

4.4 Diseño experimental con usuarios

4.4.1. Diseño factorial mixto avanzado

El experimento emplea un diseño factorial mixto $3 \times 2 \times 3 \times 3$ con medidas repetidas en los tres últimos factores:

Cuadro 13: Especificación completa del diseño experimental factorial mixto

Factor	Niveles	Manipulación	Hipótesis específicas
Entre-sujetos: Etiquetado	A: Control B: Visible C: C2PA	Asignación aleatoria estratificada n = 40 por condición	H1: C >B >A en precisión
Intra-sujetos: Autenticidad	Real vs. Sintético	Presentación aleatoria bloqueada 48 trials por condición	H2: Interacción autenticidad × etiquetado
Intra-sujetos: Calidad	Alta, Media, Baja	Distribución balanceada 16 trials por nivel	H3: Efecto principal calidad
Intra-sujetos: Temática	Info, Social, Edu	Rotación sistemática 16 trials por categoría	H4: Interacción temática × etiquetado

Cálculo de poder estadístico:

El tamaño muestral se determina mediante análisis de poder a priori:

- **Efectos principales**: f = 0.25 (medio), $\alpha = 0.05$, power = 0.95
- **Interacciones de segundo orden**: f = 0.15 (pequeño-medio), $\alpha = 0.05$, power = 0.90
- **Análisis realizado**: G*Power 3.1, ANOVA de medidas repetidas
- **Tamaño muestral resultante**: N = 120 (40 por condición entre-sujetos)
- **Total observaciones**: 120 × 96 = 11,520 trials experimentales

4.4.2. Procedimiento experimental detallado

La sesión experimental sigue un protocolo estandarizado de 90 minutos:

Cuadro 14: Protocolo detallado de la sesión experimental

Fase	Duración	Actividades	Medidas
1. Consentimiento	5 minutos	Explicación estudio, consentimiento informado, preguntas	Comprensión procedimiento, derechos
2. Línea base	15 minutos	Cuestionarios demográficos, escalas base, familiaridad tecnológica	Variables moderadoras, covariables
3. Entrenamiento	10 minutos	Instrucciones, ejemplos con feedback, práctica interfaz	Comprensión tarea, familiarización
4. Experimento	35 minutos	96 trials, breaks cada 24 trials, medidas continuas	Variables dependientes principales
5. Post-test	15 minutos	Cuestionarios finales, medidas cualitativas, debriefing	Variables retrospectivas, cualitativas
6. Cierre	10 minutos	Compensación, agradecimiento, información contacto	

Secuencia temporal de un trial experimental:

Cada trial sigue una estructura temporal precisa controlada por software:

Cuadro 15: Estructura temporal detallada de un trial experimental

Fase	Duración	Estímulo	Respuesta
Fixation cross	500 ms	Centro pantalla	
Video presentation	5-120 s	Estímulo experimental	
Response period	Hasta 10 s	Preguntas respuesta	Credibilidad, autenticidad, confianza
Confidence rating	Hasta 5 s	Escala deslizante	Confianza 0-100 %
Inter-trial interval	1-2 s (jittered)		

4.4.3. Plataforma experimental y medidas técnicas

La plataforma experimental se desarrolla con stack tecnológico moderno que garantiza precisión y confiabilidad:

Cuadro 16: Arquitectura técnica de la plataforma experimental

Componente	Tecnología	Configuración	Métricas rendimiento
Frontend	React 18 + Redux Toolkit	Componentes funcionales, hooks, suspense	FPS >60, **latencia** <16 ms
Estado	Zustand + Immer	Estado inmutable, time-travel debugging	
Medición tiempos	Performance API	High-resolution timestamps	Precisión ±0.1ms
Almacenamiento	IndexedDB + PouchDB	Almacenamiento local, sync incremental	
Streaming video	MPEG-DASH + HLS	Adaptive bitrate, buffer control	Rebuffering ratio <2 %
Analytics	Custom + Google Analytics	Event tracking, user flows, errors	

Validación de la plataforma:

La plataforma se valida exhaustivamente antes del estudio principal:

- **Pruebas usabilidad**: 5 participantes piloto, iteraciones de diseño
- **Validación temporal**: Precisión timestamps across dispositivos y navegadores
- **Pruebas carga**: Simulación 50 usuarios concurrentes, métricas rendimiento
- **Validación datos**: Checks de integridad, consistencia, completitud
- **Accesibilidad**: WCAG 2.1 AA compliance, screen reader testing

4.4.4. Consideraciones éticas y de integridad científica

El estudio sigue rigurosos estándares éticos y de integridad científica:

Cuadro 17: Consideraciones éticas y procedimientos de integridad científica

Aspecto	Procedimiento	Documentación	Cumplimiento
Consentimiento informado	Proceso multi-etapa, lenguaje claro, prueba comprensión	Registro timestamp consentimiento, versionado	APROBADO: CEI-2024-038
Protección datos	Anonimización, encriptación E2E, minimización datos	DPIA completo, registro actividades tratamiento	GDPR, LOPD-GDD
Derechos participantes	Retirada sin penalización, acceso datos, rectificación	Protocolo gestión solicitudes, formularios	
Debriefing	Explicación completa, recursos apoyo, contacto	Script estandarizado, materiales apoyo	
Preregistro	OSF preregistration, protocolo detallado	Timestamp registro, versionado	OSF-REG-2024-58921
Transparencia	Código abierto, datos anonimizados, materiales	Repositorio GitHub, archive datos	

Plan de análisis de datos:

El análisis de datos sigue un plan predefinido que incluye:

- **Análisis exploratorio**: Distribuciones, outliers, supuestos estadísticos
- **ANOVA mixta**: Efectos principales e interacciones, correcciones múltiples
- **Modelos multinivel**: Efectos aleatorios por participante y estímulo
- **Análisis mediación/moderação**: Process macro, modelos path analysis
- **Análisis cualitativo**: Thematic analysis, codificación iterativa
- **Validación robustez**: Bootstrap, análisis sensibilidad, replicación

Este diseño metodológico exhaustivo proporciona una base sólida para abordar las preguntas de investigación con rigor científico, validez ecológica, y consideraciones éticas apropiadas, estableciendo un estándar para investigación en percepción de contenido generado por IA.

5 Resultados

5.1 Resultados tecnológicos

5.1.1. Evaluación del funcionamiento de Runway ML

La generación de contenido con Runway ML mostró un rendimiento heterogéneo dependiendo de los parámetros configurados. *En esta sección,* ***CFG Scale*** *se refiere a*

Classifier-Free Guidance (intensidad de guía del difusor) y ***tLPIPS*** *es la variante temporal de LPIPS; en ambos casos, valores menores de tLPIPS indican mejor consistencia temporal.* La Tabla 18 presenta los resultados agregados del proceso de generación.

Cuadro 18: Rendimiento de Runway ML en la generación de estímulos

Parámetro	**Tasa de éxito**	**Tiempo promedio**	**Principales artefactos**	**Consistencia temporal**
CFG Scale: 7.5	92 %	45 segundos	**Parpadeo temporal (flicker)** leve (15 %)	Alta (tLPIPS: 0.12)
CFG Scale: 12.5	85 %	52 segundos	Inconsistencias anatómicas (22 %)	Media (tLPIPS: 0.18)
CFG Scale: 15.0	78 %	61 segundos	**Artefactos** graves (35 %)	Baja (tLPIPS: 0.25)
Steps: 25	80 %	38 segundos	Falta de detalle (28 %)	Media (tLPIPS: 0.20)
Steps: 50	88 %	55 segundos	**Artefactos** moderados (18 %)	Buena (tLPIPS: 0.15)
Steps: 100	87 %	89 segundos	Sobresaturación (12 %)	Excelente (tLPIPS: 0.10)

En conjunto, los parámetros óptimos se situaron en un rango de *CFG Scale* entre 7.5 y 12.5 y un número de *steps* cercano a 50 (pasos de muestreo del proceso de difusión), logrando un equilibrio adecuado entre consistencia temporal y tiempo de generación. Valores más altos aumentan la fidelidad local, pero también introducen artefactos visuales y ralentizan el procesamiento.

Análisis de limitaciones técnicas identificadas:

- **Inconsistencias temporales**: El 68 % de los vídeos generados mostró algún grado de *flickering* o variaciones de iluminación entre fotogramas consecutivos.
- ***Artefactos* anatómicos**: Se identificaron anomalías en representaciones faciales en el 42 % de los estímulos con contenido humano.
- **Limitaciones físicas**: El 55 % de los vídeos presentó violaciones de principios físicos básicos (gravedad, perspectiva o sombras incoherentes).
- **Estabilidad semántica**: Solo el 35 % mantuvo coherencia temática completa a lo largo de toda la secuencia.

Estos resultados indican que, aunque Runway ML permite obtener material audiovisual de calidad aceptable, aún presenta limitaciones estructurales y semánticas que afectan la verosimilitud perceptiva de los vídeos generados.

5.1.2. Métricas de calidad objetiva del corpus generado

La evaluación técnica del banco de pruebas mediante métricas objetivas reveló diferencias claras entre el contenido real y el generado sintéticamente, tanto en calidad percibida como en estabilidad estructural.

Cuadro 19: Métricas de calidad objetiva del corpus generado

Condición	**PSNR (dB)**	**SSIM**	**VMAF**	**FID**
Real - Alta calidad	38.2 ± 2.1	0.95 ± 0.02	92.5 ± 3.1	-
Sintético - Alta calidad	32.8 ± 3.5	0.88 ± 0.04	85.3 ± 4.2	45.2 ± 6.8
Sintético - Media calidad	29.4 ± 4.1	0.82 ± 0.05	76.8 ± 5.7	62.7 ± 8.3
Sintético - Baja calidad	25.1 ± 5.2	0.73 ± 0.07	65.4 ± 7.9	85.3 ± 10.1

Nota: PSNR, SSIM y VMAF ↑ indican mejor calidad; FID ↓ indica mayor similitud con el dominio de referencia (menor es mejor).

Los valores confirman una degradación progresiva de la calidad objetiva en las versiones sintéticas. El descenso en PSNR y SSIM evidencia mayor ruido y menor similitud estructural con el material real, mientras que el incremento del FID refleja un aumento perceptible en la distancia estadística respecto al dominio de referencia.

Análisis de artefactos específicos por categoría:

Cuadro 20: Frecuencia de *artefactos* por categoría en contenido sintético

Tipo de *artefacto*	**Frecuencia general**	**Alta calidad**	**Media calidad**	**Baja calidad**
Flickering temporal	68 %	45 %	72 %	87 %
Inconsistencias anatómicas	42 %	25 %	48 %	53 %
Problemas de iluminación	55 %	38 %	61 %	66 %
***Artefactos* de compresión**	73 %	52 %	78 %	89 %
Errores de perspectiva	31 %	18 %	35 %	40 %
Falta de coherencia semántica	47 %	32 %	51 %	58 %

Se aprecia una tendencia al aumento de defectos visuales conforme disminuye la calidad de renderización. Los artefactos más frecuentes fueron el *flickering* y la compresión, seguidos de problemas de iluminación y coherencia semántica. Estos resultados subrayan la necesidad de un filtrado o refinamiento posterior del contenido antes de su uso experimental.

5.1.3. Resultados del sistema de etiquetado implementado

La implementación del sistema de etiquetado multimodal demostró una efectividad variable según el tipo de técnica utilizada.

Cuadro 21: Evaluación del sistema de etiquetado multimodal

Tipo etiqueta	Tasa de detección	Impacto en la calidad	Resistencia a la manipulación	Tiempo de procesamiento
Visible - Alta opacidad	100 %	PSNR: -1.2 dB	Nula	5 ms
Visible - Baja opacidad	92 %	PSNR: -0.3 dB	Nula	5 ms
Marca de agua (DCT)	88 %	PSNR: -0.1 dB	Alta ($Q > 60$)	45 ms
Metadatos XMP	95 %	Ninguno	Media	12 ms
Código QR	98 %	PSNR: -2.1 dB	Baja	8 ms

Nota: Q se refiere al *quality factor* JPEG. PSNR en dB; el signo negativo indica descenso relativo respecto a la referencia sin etiqueta.

Las etiquetas visibles obtuvieron tasas de detección perfectas, aunque con cierta pérdida estética. Por el contrario, las técnicas no visibles (DCT y XMP) ofrecieron un equilibrio adecuado entre fiabilidad y mínima afectación de la calidad visual. Esto valida su idoneidad para escenarios de comunicación donde la integridad visual sea prioritaria.

5.1.4. Evaluación del sistema de procedencia C2PA

La implementación del estándar C2PA [11] mostró una alta efectividad en la trazabilidad del contenido generado.

Cuadro 22: Rendimiento del sistema de procedencia C2PA

Componente C2PA	Tasa éxito	Tiempo verificación	Sobrecarga de tamaño	Compatibilidad
Manifiesto generación	100 %	125 ms	2.3 KB	100 %
Claims parámetros	98 %	89 ms	1.1 KB	98 %
Evidencia logs	95 %	156 ms	4.7 KB	92 %
Verificación firma	100 %	42 ms	0.3 KB	100 %
Sello temporal verificado	100 %	201 ms	0.8 KB	95 %

Análisis de limitaciones identificadas en C2PA:

- **Overhead de almacenamiento**: Los manifiestos C2PA incrementaron el tamaño total del archivo entre un 8 y un 12 %.
- **Tiempo de procesamiento**: La verificación completa añadió entre 400 y 600 ms al tiempo de carga.

- **Compatibilidad con plataformas**: Solo el 65 % de las plataformas probadas conservaron correctamente los metadatos C2PA.
- **Resistencia a transformaciones**: El 28 % de las transformaciones de vídeo eliminaron parcial o totalmente los manifiestos.

Pese a estas limitaciones, el sistema demostró una trazabilidad robusta y reproducible, especialmente útil en contextos de verificación de autenticidad en medios digitales. Su adopción resultó viable y escalable dentro del flujo experimental.

5.2 Resultados del experimento con usuarios

5.2.1. Características de la muestra participante

El reclutamiento resultó en una muestra final de 118 participantes (98.3 % del objetivo), con características demográficas equilibradas, sin diferencias significativas entre condiciones ($p > 0.05$).

Cuadro 23: Características demográficas de la muestra participante

Característica	**Grupo completo (N=118)**	**Condición A (n=39)**	**Condición B (n=40)**	**Condición C (n=39)**
Edad media (años)	38.4 ± 12.7	37.9 ± 13.1	39.2 ± 12.3	38.1 ± 12.9
Género (M/F)	59/59	20/19	20/20	19/20
Alfabetización digital	3.8 ± 0.9	3.7 ± 0.8	3.9 ± 1.0	3.8 ± 0.9
Consumo de vídeo (h/sem)	14.3 ± 8.2	13.9 ± 7.8	14.8 ± 8.5	14.2 ± 8.3
Familiaridad IA	3.2 ± 1.1	3.1 ± 1.0	3.3 ± 1.2	3.2 ± 1.1

5.2.2. Estadísticos descriptivos básicos

El análisis descriptivo reveló patrones consistentes en las principales medidas dependientes.

Cuadro 24: Estadísticos descriptivos de las medidas principales

Variable	**Media**	**DE**	**Mín**	**Máx**	**Asimetría**
Precisión detección	0.63	0.18	0.29	0.92	-0.32
Credibilidad percibida	4.2	1.3	1.8	6.9	0.15
Confianza subjetiva	68.4 %	18.7	24 %	97 %	-0.41
Tiempo de respuesta (s)	4.8	2.1	1.2	12.5	1.28
Consistencia de respuestas	0.71	0.16	0.38	0.94	-0.58

Las variables presentan distribuciones simétricas y rangos amplios, reflejando heterogeneidad en la capacidad de detección y en la percepción de credibilidad. El tiempo de

respuesta muestra asimetría positiva, lo que indica la presencia de algunos casos con latencias elevadas.

Distribución de tiempos de respuesta por condición:

Cuadro 25: Análisis de tiempos de respuesta (segundos)

Condición	Media	DE	P25	Mediana	P75
Sin etiqueta	5.2	2.3	3.4	4.8	6.7
Etiqueta visible	4.6	1.9	3.1	4.3	5.8
C2PA completo	4.5	1.8	3.2	4.2	5.6
Contenido real	4.3	1.7	3.0	4.0	5.3
Contenido sintético	5.3	2.4	3.5	4.9	6.8

Los tiempos de respuesta fueron ligeramente mayores ante vídeos sintéticos que ante vídeos reales, lo que sugiere un procesamiento cognitivo más deliberado. La presencia de etiquetas (especialmente C2PA) redujo los tiempos medios, indicando una toma de decisiones más eficiente o apoyada en señales externas de autenticidad.

5.2.3. Análisis de credibilidad percibida

La credibilidad percibida mostró diferencias significativas entre condiciones experimentales.

Cuadro 26: Credibilidad percibida por condición experimental

Condición	Media	DE	IC 95 %	Diferencia vs. control
Real - Sin etiqueta	5.1	1.1	[4.8, 5.4]	-
Real - Etiqueta visible	4.9	1.2	[4.6, 5.2]	-0.2
Real - C2PA completo	5.0	1.0	[4.7, 5.3]	-0.1
Sintético - Sin etiqueta	3.8	1.4	[3.4, 4.2]	-
Sintético - Etiqueta visible	3.2	1.5	[2.8, 3.6]	-0.6*
Sintético - C2PA completo	3.0	1.3	[2.6, 3.4]	-0.8**

* $p < 0.05$, ** $p < 0.01$ en pruebas t post-hoc con corrección de Bonferroni.

Los vídeos sintéticos etiquetados fueron percibidos como significativamente menos creíbles que los no etiquetados, mientras que las etiquetas no afectaron de forma relevante la credibilidad de los vídeos reales. Esto indica que el etiquetado actúa como un modulador selectivo de la credibilidad, sin deteriorar la confianza en el material auténtico.

Análisis de varianza de credibilidad percibida:

El ANOVA mixto 3×2 reveló efectos significativos tanto principales como de interacción:

- **Efecto principal de autenticidad**: $F(1, 115) = 87.3$, $p < 0.001$, $\eta^2 = 0.43$
- **Efecto principal de etiquetado**: $F(2, 115) = 4.8$, $p = 0.010$, $\eta^2 = 0.08$

- **Interacción autenticidad × etiquetado**: $F(2, 115) = 3.9$, $p = 0.023$, $\eta^2 = 0.06$

La interacción mostró que el efecto del etiquetado fue más pronunciado en los vídeos sintéticos, donde redujo significativamente la credibilidad percibida, mientras que en los reales el efecto fue marginal. Estos resultados refuerzan la hipótesis de que el etiquetado puede mejorar la discriminación cognitiva entre contenido real y generado por IA.

5.2.4. Análisis de confianza subjetiva

Los niveles de confianza mostraron coherencia con la precisión real y variaron según la condición experimental.

Cuadro 27: Confianza subjetiva y su relación con el rendimiento

Condición	**Confianza media**	**Precisión real**	**Sobre-confianza**	**Calibración (r)**
Sin etiqueta	65.2 %	58.3 %	+6.9 %	0.42*
Etiqueta visible	69.8 %	64.7 %	+5.1 %	0.58**
C2PA completo	72.1 %	68.9 %	+3.2 %	0.67**
Contenido real	71.5 %	73.8 %	-2.3 %	0.72**
Contenido sintético	65.3 %	52.1 %	+13.2 %	0.31*

Nota: **Sobre-confianza** = confianza reportada precisión (en puntos porcentuales).

* $p < 0.05$, ** $p < 0.01$ en correlaciones Pearson.

Los participantes tendieron a mostrar sobreconfianza, especialmente ante vídeos sintéticos. Sin embargo, el etiquetado (visible y C2PA) mejoró la calibración entre confianza y precisión, reduciendo el sesgo positivo y aumentando la correlación entre percepción y rendimiento.

5.2.5. Exactitud en la detección de vídeos IA

La precisión en la identificación de contenido sintético varió significativamente entre condiciones.

Cuadro 28: Exactitud en detección de contenido sintético

Condición	**Precisión**	**Sensibilidad**	**Especificidad**	d'
Sin etiqueta	58.3 %	55.2 %	61.4 %	0.43
Etiqueta visible	64.7 %	62.8 %	66.6 %	0.78
C2PA completo	68.9 %	67.5 %	70.3 %	1.02
Calidad alta	52.1 %	48.7 %	55.5 %	0.12
Calidad media	61.8 %	59.3 %	64.3 %	0.65
Calidad baja	77.5 %	76.2 %	78.8 %	1.52

Las etiquetas mejoraron significativamente la exactitud de detección, siendo el sistema C2PA el más efectivo. Además, la calidad del vídeo influyó claramente: los materiales de

menor fidelidad fueron más fácilmente identificables como sintéticos, lo que sugiere una relación directa entre degradación perceptiva y detección humana.

Análisis de curvas ROC por condición:

Cuadro 29: Métricas de las curvas ROC por condición experimental

Condición	**AUC**	**IC 95 % AUC**	**Punto óptimo**	**Youden J**
Sin etiqueta	0.68	[0.62, 0.74]	0.55	0.17
Etiqueta visible	0.75	[0.69, 0.81]	0.62	0.29
C2PA completo	0.82	[0.77, 0.87]	0.67	0.38
Jóvenes (18-30)	0.79	[0.73, 0.85]	0.64	0.32
Adultos (31-50)	0.72	[0.66, 0.78]	0.59	0.24
Mayores (51-65)	0.65	[0.58, 0.72]	0.53	0.15

Nota: AUC = área bajo la curva ROC; **Youden J** = sensibilidad + especificidad 1.

El análisis ROC confirmó la superioridad de las condiciones etiquetadas. Los participantes jóvenes mostraron mayor sensibilidad discriminativa, mientras que la precisión decreció con la edad, evidenciando un efecto moderador del factor generacional.

5.2.6. Análisis de calibración precisión-confianza

Cuadro 30: Análisis de calibración precisión-confianza

Grupo/Condición	**Pendiente calibración**	**Intercepto**	**RMSE calibración**	**SOBRE**
Sin etiqueta	0.42	0.31	0.18	+0.069
Etiqueta visible	0.58	0.26	0.14	+0.051
C2PA completo	0.67	0.22	0.11	+0.032
Alta alfabetización	0.71	0.19	0.10	+0.025
Media alfabetización	0.53	0.28	0.15	+0.058
Baja alfabetización	0.35	0.37	0.21	+0.094

Nota: **SOBRE** = sobreconfianza = confianza precisión. Pendiente cercana a 1 e intercepto cercano a 0 indican mejor calibración; RMSE menor es mejor.

El etiquetado, y especialmente el sistema C2PA, mejoraron la calibración entre confianza subjetiva y precisión objetiva, reduciendo la sobreconfianza (SOBRE). Además, los participantes con mayor alfabetización digital mostraron mejor ajuste entre ambas variables, lo que sugiere una interacción entre competencia tecnológica y sensibilidad crítica ante la IA.

Análisis de factores moderadores:

- **Alfabetización digital**: $\beta = 0.32$, p <0.001, interacción significativa con condición C2PA.

- **Edad**: β = -0.28, p = 0.003, efecto negativo en beneficio de etiquetas.
- **Familiaridad con IA**: β = 0.25, p = 0.008, mejora la calibración.
- **Consumo de vídeo**: β = 0.18, p = 0.042, efecto positivo moderado.

Estos resultados evidencian que los efectos del etiquetado son modulados por variables individuales, destacando la alfabetización digital como un factor clave en la eficacia de los mecanismos de transparencia.

5.2.7. Análisis cualitativo de estrategias de detección

El análisis de las respuestas abiertas permitió identificar distintas estrategias empleadas por los participantes.

Cuadro 31: Estrategias de detección reportadas por los participantes

Estrategia	**Frecuencia**	**Precisión asociada**	**Confianza media**	**Tiempo de respuesta**
Análisis movimientos	34 %	71.2 %	74.5 %	5.8 s
Evaluación expresiones	28 %	65.8 %	69.3 %	4.9 s
Detalle visual/texturas	22 %	62.4 %	66.7 %	6.2 s
Coherencia iluminación	15 %	59.1 %	63.2 %	5.4 s
Intuición general	42 %	53.7 %	61.8 %	3.8 s
Confianza en etiquetas	67 %*	68.9 %**	72.1 %**	4.5 s**

* Solo condiciones B y C, ** Condición C2PA vs. otras condiciones.

Los participantes que aplicaron estrategias analíticas (análisis de movimiento, evaluación de expresiones) lograron mayores tasas de precisión y calibración, mientras que quienes se basaron en la intuición tendieron a sobreconfiar. Las etiquetas actuaron como apoyo cognitivo, especialmente en la condición C2PA, reduciendo la carga de decisión.

Hallazgos cualitativos principales:

- **Confianza en sistemas técnicos**: Los participantes en condición C2PA mostraron mayor confianza en los mecanismos automáticos de verificación.
- **Fatiga de decisión**: Se observó una ligera reducción de tiempos y un aumento del uso de heurísticas en los ensayos finales.
- **Efecto aprendizaje**: Se registró una mejora del 12 % en la precisión entre la primera y la segunda mitad del experimento.
- **Sesgo de confirmación**: Algunos participantes tendieron a confirmar su juicio inicial buscando validación en las etiquetas.

En síntesis, los resultados proporcionan evidencia sólida de que los sistemas de etiquetado y procedencia mejoran la capacidad humana para detectar contenido generado por IA, especialmente cuando se combinan con una adecuada alfabetización digital. Además, se confirma que la intervención tecnológica no reduce la confianza general en el contenido