RECONSTRUCCIÓN TEMPORAL MEDIANTE CODIFICACIÓN PASIVA EN IMÁGENES TÉRMICAS USANDO MODELOS DE VISIÓN Y LENGUAJE.

LUIS MARIO TOSCANO PALOMINO
GABRIEL DAVID CASTILLO RODRIGUEZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2026

RECONSTRUCCIÓN TEMPORAL MEDIANTE CODIFICACIÓN PASIVA EN IMÁGENES TÉRMICAS USANDO MODELOS DE VISIÓN Y LENGUAJE.

LUIS MARIO TOSCANO PALOMINO
GABRIEL DAVID CASTILLO RODRIGUEZ

Trabajo de Grado para optar al título de
Ingeniero de Sistemas

Director:
Lola Xiomara Bautista Rozo
Magíster/Ph.D. en Tratamiento de Señales e Imágenes

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2026

DEDICATORIA

A Dios, por la sabiduría y fortaleza en cada etapa de este camino.

A nuestras familias, pilares fundamentales de amor y apoyo incondicional.

A nuestros amigos, por su compañía y confianza en los momentos de exigencia.

Y a todos quienes, con su apoyo, hicieron posible alcanzar este objetivo.

AGRADECIMIENTOS

Agradecemos de manera especial a los profesores Jorge Bacca y Lola Xiomara por su valiosa guía, acompañamiento y dirección a lo largo del desarrollo de este trabajo de grado. Su conocimiento, disposición y orientación fueron fundamentales para la culminación exitosa de este proyecto.

De igual forma, expresamos nuestro agradecimiento a los compañeros del grupo PIGroup, quienes contribuyeron significativamente mediante la revisión detallada del trabajo, el refinamiento de ideas y la aportación de consejos valiosos que fortalecieron la calidad del mismo.

Finalmente, agradecemos a nuestros compañeros de carrera, hoy también amigos, quienes estuvieron presentes brindando apoyo, colaboración y ayuda oportuna en los momentos en que más lo necesitamos.

CONTENIDO

pág.

LISTA DE FIGURAS

pág.

LISTA DE CUADROS

pág.

LISTA DE ANEXOS

pág.

RESUMEN

TÍTULO: RECONSTRUCCIÓN TEMPORAL MEDIANTE CODIFICACIÓN PASIVA EN IMÁGENES TÉRMICAS USANDO MODELOS DE VISIÓN Y LENGUAJE. *

AUTOR: LUIS MARIO TOSCANO PALOMINO
GABRIEL DAVID CASTILLO RODRIGUEZ **

PALABRAS CLAVE: Modelos generativos multimodales, Problemas inversos mal condicionados, Señales térmicas, Reconstrucción temporal.

DESCRIPCIÓN:

El presente trabajo estudia la reconstrucción temporal de escenas a partir de señales residuales mediante el paradigma de *Time-reversed Imaging*. Mientras que la visión artificial convencional se centra en interpretar el estado actual de una escena o anticipar eventos futuros a partir de observaciones directas, esta investigación busca inferir interacciones humanas recientes utilizando huellas térmicas que permanecen en proceso de disipación. Desde una perspectiva física, recuperar información sobre un estado pasado a partir de mediciones actuales constituye un problema inverso mal condicionado, ya que las trazas térmicas son incompletas, ruidosas y pueden corresponder a múltiples eventos diferentes.

Para abordar esta limitación, se propone un sistema multimodal que combina imágenes térmicas y RGB capturadas en el mismo instante. Ambas modalidades son analizadas por un modelo de visión y lenguaje (VLM), encargado de generar una descripción semántica de la evidencia observada. Esta representación textual se utiliza posteriormente como condición de entrada para un modelo generativo basado en difusión, el cual reconstruye una posible representación visual del estado previo de la escena.

El sistema se prueba en interacciones físicas ocurridas desde unos pocos segundos hasta dos minutos antes de la captura. Además, se estudia el efecto que tiene la pérdida gradual de información térmica sobre la reconstrucción de la escena. Los resultados obtenidos muestran que la combinación de imágenes térmicas, visión artificial y modelos generativos permite recuperar información útil sobre eventos recientes. Finalmente, se discuten las limitaciones del método y los retos de aplicarlo en escenarios reales.

---

* Trabajo de grado

** FACULTAD DE INGENIERÍAS FISICOMECÁNICAS. ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA. Director: Lola Xiomara Bautista Rozo.

ABSTRACT

TITLE: Temporal reconstruction through passive encoding in thermal images using vision-language models. [*]

AUTHOR: LUIS MARIO TOSCANO PALOMINO
GABRIEL DAVID CASTILLO RODRIGUEZ [**]

KEYWORDS: Multimodal generative models, Ill-posed inverse problems, Thermal signals, Temporal reconstruction.

DESCRIPTION:

This work explores the reconstruction of past scenes from residual thermal signals using the *Time-reversed Imaging* paradigm. While conventional computer vision mainly focuses on understanding the current state of a scene or predicting future events, this approach attempts to infer recent human interactions from thermal traces that remain after contact. From a physical perspective, recovering previous states from current observations is an ill-posed inverse problem because thermal traces are noisy, incomplete, and may correspond to different past events.

The proposed framework combines synchronized RGB and thermal images captured at the same instant. Both modalities are processed by a vision-language model (VLM), which generates a semantic description of the observed interaction. This description is then used to guide a diffusion-based generative model that reconstructs a plausible representation of the scene before the interaction occurred.

The system is evaluated on interactions that took place from a few seconds up to two minutes before image acquisition. In addition, the experiments analyze how reconstruction quality changes as thermal information fades over time due to natural cooling. Results show that combining thermal evidence with semantic guidance improves the reconstruction of recent events. Finally, the limitations of the current approach and the challenges of applying the method in real-world environments are discussed.

---

[*] Bachelor Thesis

[**] FACULTAD DE INGENIERÍAS FISICOMECÁNICAS. ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA. Advisor: Lola Xiomara Bautista Rozo.

INTRODUCCIÓN

Inferir qué ocurrió recientemente a partir de observaciones posteriores al evento constituye un problema inverso altamente mal condicionado. A partir de una única imagen RGB, el número de eventos pasados que podrían haber generado el estado observado son ilimitados. Como se ilustra en la Fig. 1, un análisis basado únicamente en información visible no permite distinguir entre escenarios fundamentalmente distintos, como una interacción breve con una superficie o un contacto prolongado con la misma. Esta limitación se vuelve crítica en escenarios reales de *observación tardía*, donde el evento de interés ya ha ocurrido y solo se dispone de sus efectos residuales.

En diversos contextos aplicados, los sistemas deben razonar sobre interacciones pasadas a partir de información incompleta o diferida. Por ejemplo, en análisis forense, los investigadores suelen llegar después de ocurrido un incidente[1]; en sistemas de seguridad, pueden detectarse estados anómalos sin haber registrado la acción que los generó[2]; en entornos colaborativos, como la interacción humano-robot, es necesario inferir manipulaciones recientes[3]. En entornos inteligentes, como hogares o sistemas asistidos, únicamente se observan consecuencias de interacciones, tales como objetos desplazados o rastros térmicos[4]. Estos escenarios evidencian la necesidad de desarrollar métodos computacionales capaces de inferir eventos recientes a partir de evidencia indirecta.

---

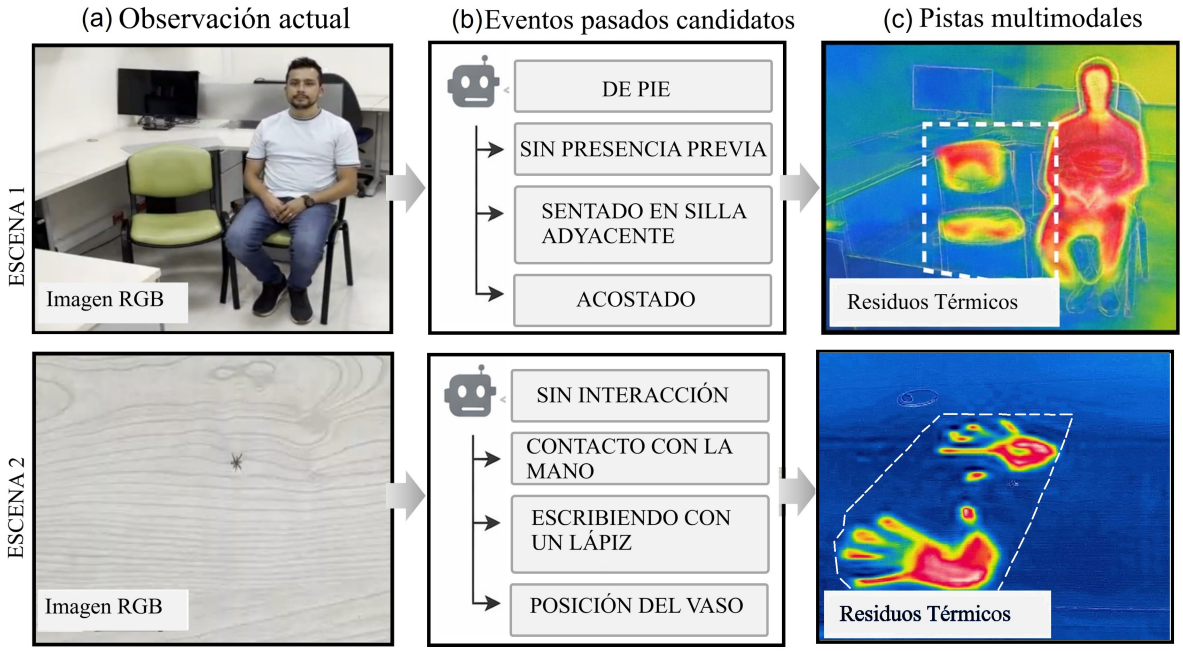
1 Ioannis KETSEKIOULAFIS *et al.* «Artificial Intelligence in Forensic Sciences: a systematic review of past and current applications and future perspectives». En: *Cureus* 16.9 (2024).

2 Arief KOESDWIADY *et al.* «Recent trends in driver safety monitoring systems: State of the art and challenges». En: *IEEE Transactions on Vehicular Technology* 66.6 (2016), pp. 4550-4563.

3 Nicole ROBINSON *et al.* «Robotic vision for human-robot interaction and collaboration: A survey and systematic review». En: *ACM Transactions on Human-Robot Interaction* 12.1 (2023).

4 M. L. CÓRDOBA-TLAXCALTECO y Edgard BENÍTEZ-GUERRERO. «Human event recognition in smart classrooms using computer vision: a systematic literature review». En: *Programming and Computer Software* 49.8 (2023), pp. 625-642.

Figura 1. **Problema mal condicionado en la inferencia temporal inversa y rol de las huellas térmicas.** A partir de una observación RGB actual, el conjunto de posibles eventos pasados es inherentemente ambiguo, ya que múltiples interacciones pueden producir estados visuales similares. La incorporación de información térmica permite capturar huellas físicas residuales de contacto reciente, proporcionando restricciones adicionales que reducen la ambigüedad del problema.



Fuente: Elaboración propia.

Los métodos tradicionales de razonamiento temporal en visión por computadora, como la interpolación y extrapolación de video[5], operan en la dirección directa del tiempo, prediciendo estados futuros a partir de observaciones pasadas. Aunque estos enfoques han demostrado la capacidad de aprender dinámicas temporales, dependen exclusivamente de patrones estadísticos en los datos visuales. En consecuencia, la inferencia del pasado

[5] Hang GAO *et al.* «Disentangling propagation and generation for video prediction». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 9006-9015; Ziwei LIU *et al.* «Video frame synthesis using deep voxel flow». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2017, pp. 4463-4471; Wei YU *et al.* «Crevnet: Conditionally reversible video prediction». En: *arXiv preprint arXiv:1910.11577* (2019); William LOTTER, Gabriel KREIMAN y David COX. «Deep predictive coding networks for video prediction and unsupervised learning». En: *arXiv preprint arXiv:1605.08104* (2016).

basada únicamente en información RGB resulta estructuralmente indeterminada por la carencia de evidencia física latente.

Sin embargo, los entornos reales conservan una forma de memoria física. Las interacciones entre humanos y el entorno dejan huellas residuales que persisten durante intervalos cortos de tiempo tras el contacto, particularmente en forma de señales térmicas. Estas huellas codifican patrones espaciales asociados a interacciones recientes y evolucionan de acuerdo con procesos físicos como la disipación de calor[6]. A diferencia de la información visible, estas señales proporcionan evidencia medible del pasado reciente, y han demostrado ser útiles para inferir eventos recientes[7], permitiendo restringir el espacio de soluciones del problema inverso.

A partir de esta base, este trabajo plantea la exploración de un enfoque de reconstrucción temporal basado en señales térmicas residuales. La propuesta consiste en integrar información térmica con una imagen RGB actual y emplear modelos de visión y lenguaje (VLM) para extraer representaciones semánticas estructuradas de las huellas detectadas[8]. Estas representaciones son posteriormente utilizadas como condiciones en un modelo generativo basado en difusión[9], con el objetivo de sintetizar reconstrucciones visuales plausibles del estado previo de la escena.

El objetivo de este estudio es analizar la viabilidad de este enfoque multimodal, evaluar su estabilidad frente a la degradación temporal de las señales térmicas y estudiar los límites de recuperación de información en ventanas de tiempo cortas posteriores a una

---

6 Michael VOLLMER y Klaus-Peter MÖLLMANN. *Infrared thermal imaging: fundamentals, research and applications*. John Wiley & Sons, 2018.

7 Daniela CARDONE y Arcangelo MERLA. «New frontiers for applications of thermal infrared imaging devices». En: *Sensors* 17.5 (2017), pág. 1042; Zitian TANG *et al.* «What happened 3 seconds ago? Inferring the past with thermal imaging». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 17111-17120.

8 Alec RADFORD *et al.* «Learning transferable visual models from natural language supervision». En: *Proceedings of the International Conference on Machine Learning (ICML)*. 2021, pp. 8748-8763; Peng WANG *et al.* «Qwen2-VL: Enhancing vision-language model's perception». En: *arXiv preprint arXiv:2409.12191* (2024); Haotian LIU *et al.* «Improved baselines with visual instruction tuning». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 26296-26306.

9 Robin ROMBACH *et al.* «High-resolution image synthesis with latent diffusion models». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10684-10695.

interacción. En particular, se determinó en qué medida es posible inferir eventos ocurridos en intervalos del orden de segundos hasta aproximadamente dos minutos, dependiendo de las condiciones del entorno y la calidad de las señales.

De este modo, el trabajo contribuye a la comprensión del problema de reconstrucción temporal a partir de señales físicas residuales, integrando conceptos de visión por computadora, modelado físico y modelos generativos. También se explora la posible aplicación del enfoque en análisis forense y seguridad, así como las limitaciones del problema y la necesidad de validarlo en escenarios reales.

1. PLANTEAMIENTO Y JUSTIFICACION DEL PROBLEMA

La reconstrucción temporal de una escena puede entenderse como el proceso de recuperar un estado pasado a partir de huellas o señales residuales. Desde el punto de vista matemático, este problema puede describirse como un problema inverso mal condicionado[10]. La principal dificultad es que el evento original no queda almacenado directamente, sino solo mediante huellas residuales que se debilitan con el tiempo. A medida que estas señales desaparecen, resulta más difícil interpretar lo que ocurrió en la escena.

Aun así, las imágenes térmicas conservan cierta información sobre interacciones recientes, incluyendo la zona de contacto, la forma aproximada de la huella y la duración de la interacción física. Aunque no contienen información visual explícita, muestran patrones residuales que actúan como pistas físicas y temporales[11] que pueden ser aprovechadas computacionalmente para mitigar la ambigüedad del problema inverso. Abordar esta reconstrucción tiene un valor práctico significativo en áreas como análisis forense, seguridad y reconstrucción de escenas sin registros visuales directos[12]. Sin embargo, el campo del estudio de la inversión temporal aplicada directamente a señales térmicas aún está poco explorado. Esta tesis explora esta temática mediante un estudio centrado en señales térmicas que analice formalmente la estabilidad de la reconstrucción y la información recuperable. Las limitaciones que esta tesis abordará se centran en: (1) La carencia metodológica para formalizar la inversión temporal y caracterizar la contribución de cada componente del sistema a la recuperación del estado. (2) La falta de un estudio formal que analice la estabilidad de la reconstrucción y la información recuperable a par-

---

[10] Mario BERTERO y Patrizia BOCCACCI. *Introduction to Inverse Problems in Imaging*. Institute of Physics Publishing, 1998.

[11] M. MAZUR-MILECKA *et al.* «Detection and Model of Thermal Traces Left after Aggressive Behavior of Laboratory Rodents». En: *Applied Sciences* 11.14 (2021), pág. 6644.

[12] Yomna ABDELRAHMAN *et al.* «Stay Cool! Understanding Thermal Attacks on Mobile-based User Authentication». En: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI)*. 2017, pp. 3751-3763.

tir de la señal térmica, a pesar de la existencia de flujos de trabajo prometedores[13]. Por lo tanto, el trabajo se delimita a un tipo de señal específica, un flujo de trabajo definido y un conjunto de modelos manejables, para mantener el equilibrio entre rigor científico y factibilidad técnica. Recientemente, la investigación en visión térmica sugiere que las imágenes infrarrojas pueden contener trazas físicas de interacciones recientes[14] (p. ej., calor residual en superficies), lo cual aporta evidencia adicional traducido en más datos frente a la ambigüedad presentada de inferir el pasado desde una sola observación RGB. En esta línea, resulta razonable plantear que modelos multimodales (incluyendo VLM) y procesos generativos (p. ej., difusión como mecanismo de síntesis condicionada) podrían servir como piezas para explorar reconstrucciones visuales tentativas de estados previos, a partir de descriptores inferidos de firmas térmicas y del contexto. Más que fijar un horizonte temporal único, este trabajo propone estudiar ventanas temporales cortas (del orden de segundos a minutos) y caracterizar empíricamente sus límites, estabilidad e incertidumbre según el escenario y las condiciones físicas de la escena.

A partir de lo anterior, surge la pregunta de investigación: ¿Cómo diseñar un algoritmo de reconstrucción temporal que, utilizando una huella térmica (codificación pasiva) y una imagen RGB actual, permita identificar los objetos en contacto y generar una reconstrucción visual coherente del estado original de la escena RGB antes del contacto mediante un modelo de visión y lenguaje?

---

[13] TANG *et al.* Ver n. 7.

[14] TANG *et al.* Ver n. 7.

2. OBJETIVOS

Objetivo general

- Diseñar e implementar un algoritmo de reconstrucción temporal que, a partir de una huella térmica y una imagen RGB actual de la escena, permita identificar los objetos en contacto y reconstruir una imagen RGB plausible del estado previo de la escena mediante un modelo de visión y lenguaje.

Objetivos específicos

1. Construir un conjunto de datos multimodal que comprenda pares de imágenes RGB y térmicas alineadas, capturadas en escenarios controlados (sentarse, tocar, apoyarse) con intervalos de degradación térmica de 0, 5, 15, 30 y 120 segundos para servir como base de entrenamiento, validación y testeo.
2. Diseñar e implementar un modelo descriptor que integre las huellas térmicas y la imagen RGB actual para la identificación de los objetos en contacto y la generación de la representación semántica del evento previo.
3. Integrar el modelo descriptor con sistemas generativos para reconstruir el estado previo de la escena orientado por la representación semántica, considerando la evolución temporal de la huella térmica.
4. Evaluar el desempeño del sistema propuesto mediante métricas de calidad visual, coherencia semántica y estabilidad en la reconstrucción del pasado, comparando los resultados obtenidos con registros reales de los eventos.

3. ESTADO DEL ARTE Y MARCO TEÓRICO

La reconstrucción de estados pasados a partir de información incompleta ha sido estudiada desde diferentes áreas del conocimiento. Los enfoques iniciales se fundamentaron principalmente en modelos físicos y formulaciones de problemas inversos, mientras que los avances recientes en inteligencia artificial han permitido incorporar modelos generativos y sistemas multimodales para abordar este desafío desde una perspectiva basada en datos.

3.1. PROBLEMAS INVERSOS Y REGULARIZACIÓN EN TERMOGRAFÍA

El estudio de las señales térmicas transitorias se encuentra estrechamente relacionado con los problemas inversos asociados a la difusión del calor. En este tipo de problemas, el objetivo consiste en estimar condiciones pasadas a partir de observaciones realizadas en un instante posterior. Sin embargo, la información térmica disponible se degrada progresivamente con el tiempo, lo que dificulta recuperar de forma única el estado que originó la observación.

Vogel et al.[15] señalan que esta dificultad surge debido a la pérdida irreversible de información durante el proceso de difusión térmica. Como consecuencia, diferentes estados iniciales pueden generar distribuciones de temperatura muy similares después de un intervalo de tiempo Δt. Para enfrentar este problema, diversos enfoques han incorporado técnicas de regularización basadas en conocimiento previo de los materiales y del sistema observado. No obstante, la efectividad de estas estrategias suele depender de condiciones de adquisición controladas y de parámetros físicos conocidos con suficiente precisión.

Una de las primeras aproximaciones experimentales orientadas al aprovechamiento de huellas térmicas fue propuesta por Larson et al.[16] mediante el sistema HeatWave. Este

---

[15] Curtis R. VOGEL. *Computational Methods for Inverse Problems*. Society for Industrial y Applied Mathematics, 2002. DOI: 10.1137/1.9780898717570.

[16] Eric LARSON *et al.* «HeatWave: Thermal imaging for surface user interaction». En: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2011. DOI: 10.1145/1978942.

trabajo demostró que las variaciones térmicas dejadas por una persona sobre una superficie pueden utilizarse para detectar interacciones recientes y distinguir eventos de contacto y aproximación. A pesar de estos resultados, el enfoque se centra en la identificación de la interacción ocurrida y no en la reconstrucción visual del estado previo de la escena.

3.2. SÍNTESIS Y PREDICCIÓN TEMPORAL EN VIDEO

Para comprender la dinámica temporal de las escenas, la literatura ha desarrollado diversos métodos de predicción. Uno de los trabajos pioneros es *Video frame synthesis using deep voxel flow*[17], el cual propone el uso de flujo de vóxeles profundos para sintetizar fotogramas intermedios a partir de secuencias RGB.

Aportes clave:

- Introducción del *Deep Voxel Flow* para modelar correspondencias espacio-temporales.
- Síntesis de frames intermedios sin necesidad de reconstrucción explícita de la escena.

Bajo un paradigma no supervisado, el trabajo *Deep Predictive Coding Networks for Video Prediction*[18] introduce una arquitectura que aprende a predecir fotogramas futuros utilizando el error de predicción como señal de entrenamiento.

Aportes clave:

- Formulación de aprendizaje no supervisado basado en predicción temporal.
- Aprendizaje de representaciones latentes dinámicas sin necesidad de etiquetas.

En el ámbito específico de la inferencia retrospectiva, *CrevNet: Conditionally reversible video prediction*[19] introduce arquitecturas basadas en redes neuronales reversibles.

1979317.

[17] Zhengqiang LIU *et al.* «Video frame synthesis using deep voxel flow». En: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[18] LOTTER, KREIMAN y COX, ver n. 5.

[19] YU *et al.* Ver n. 5.

Aportes clave:

- Uso de arquitecturas reversibles para preservar la información temporal.
- Capacidad de inferir tanto estados futuros como pasados minimizando la pérdida de información.

3.3. MODELOS GENERATIVOS BASADOS EN DIFUSIÓN Y CONDICIONAMIENTO

El avance reciente de la inteligencia artificial ha permitido abordar la reconstrucción desde la generación de imágenes. El trabajo *High-resolution image synthesis with latent diffusion models*[20] introdujo los Modelos de Difusión Latente (LDM), trasladando el proceso de difusión a un espacio latente.

Aportes clave:

- Reducción del costo computacional mediante modelado latente.
- Base arquitectónica para modelos generativos modernos de alta resolución.

La capacidad de guiar estos modelos fue expandida por Zhang et al.[21] mediante ControlNet, una arquitectura que incorpora señales de control adicionales (mapas de profundidad, bordes) a modelos de difusión preentrenados. Este principio es fundamental para la presente investigación, ya que establece que señales no puramente visuales (como las distribuciones térmicas) pueden ser utilizadas para condicionar la reconstrucción de estados visuales coherentes.

3.4. DESCRIPTORES SEMÁNTICOS Y MODELOS DE LENGUAJE (LLM / VLM)

La reconstrucción temporal propuesta en este trabajo requiere transformar la información observada en representaciones semánticas que puedan ser utilizadas por modelos generativos. Este tipo de representación ha cobrado relevancia con el desarrollo de los

---

[20] ROMBACH *et al.* Ver n. 9.

[21] Lingzhi ZHANG, Aditya RAO y Maneesh AGRAWALA. «Adding Conditional Control to Text-to-Image Diffusion Models». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. DOI: 10.1109/ICCV51070.2023.00355.

modelos de lenguaje de gran escala (LLMs), capaces de capturar relaciones semánticas complejas a partir de grandes volúmenes de datos textuales. Entre los trabajos más influyentes se encuentra GPT-3, presentado por Brown et al.[22], que demostró capacidades de aprendizaje en contexto sin necesidad de entrenamiento específico para cada tarea.
La incorporación de información visual dio lugar a los modelos de visión y lenguaje (VLMs). Dentro de esta línea, Li et al.[23] propusieron BLIP, un modelo diseñado para relacionar imágenes y texto dentro de una misma representación multimodal. Este enfoque permitió mejorar tareas como la descripción automática de imágenes y la comprensión visual guiada por lenguaje.

3.5. EDICIÓN E INTEGRACIÓN MULTIMODAL GUIADA POR INSTRUCCIONES

La convergencia entre lenguaje y visión ha impulsado la edición condicionada. Saharia et al.[24] con su modelo Imagen evidenciaron la fuerte relación entre las representaciones lingüísticas complejas y la generación de imágenes de alta fidelidad.
Avanzando en el control de la edición, *Prompt-to-Prompt image editing with cross-attention control*[25] introdujo un enfoque basado en difusión para modificar contenido visual de manera precisa.

Aportes clave:

- Uso de atención cruzada (*cross-attention*) para controlar transformaciones visuales.
- Relación directa y localizable entre descriptores semánticos y cambios en la imagen.

Un paso más allá en esta integración multimodal es InstructPix2Pix, propuesto por Brooks

---

[22] Tom B. BROWN *et al.* «Language Models are Few-Shot Learners». En: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[23] Junnan LI *et al.* «BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation». En: *Proceedings of the International Conference on Machine Learning (ICML)*. 2022.

[24] Chitwan SAHARIA *et al.* «Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding». En: *arXiv preprint arXiv:2205.11487* (2022).

[25] Amir HERTZ *et al.* «Prompt-to-Prompt Image Editing with Cross-Attention Control». En: *arXiv preprint arXiv:2208.01626* (2022).

et al.[26], el cual utiliza modelos de difusión guiados directamente por instrucciones para realizar modificaciones estructurales coherentes, sugiriendo que la inferencia del estado pasado depende no solo de las señales físicas, sino de la interpretación semántica del evento.

3.6. MÉTRICAS DE EVALUACIÓN EN RECONSTRUCCIÓN MULTIMODAL

La evaluación de modelos generativos suele apoyarse en diferentes métricas, ya que cada una describe aspectos particulares de la calidad de una reconstrucción. Flores y González[27] señalan que PSNR (Peak Signal-to-Noise Ratio) y SSIM (Structural Similarity Index) permiten medir la similitud entre una imagen reconstruida y su referencia en términos de intensidad y estructura. Sin embargo, estas métricas no siempre reflejan las diferencias perceptuales que resultan evidentes para un observador humano. Por esta razón, es común complementar el análisis con LPIPS (Learned Perceptual Image Patch Similarity), una métrica basada en características extraídas por redes neuronales profundas que permite estimar la similitud perceptual entre imágenes.

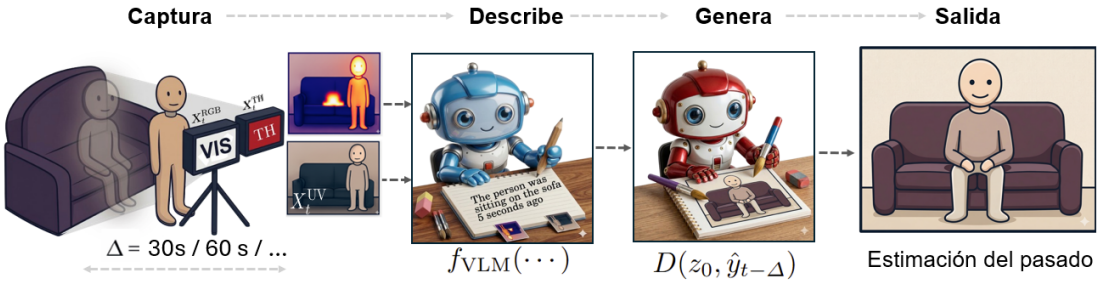
---

[26] Tom BROOKS, Anna HOLYNSKI y Alex A. EFROS. «InstructPix2Pix: Learning to follow image editing instructions». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 18392-18402. DOI: 10.1109/CVPR52729.2023.01764.

[27] Carlos FLORES y Santiago GONZÁLEZ. «Evaluación y Comparación de Métricas Objetivas PSNR, SSIM y LPIPS para el Análisis de Calidad de Video». En: *Revista Tecnológica - ESPOL* 37.E1 (2025), pág. 1317. DOI: 10.37815/rte.v37nE1.1317.

4. METODOLOGÍA PROPUESTA

Figura 2. **Arquitectura general del sistema propuesto.** Las imágenes RGB y térmicas capturadas en el instante actual se utilizan como entrada del sistema. Un modelo visión-lenguaje (VLM) genera una descripción del evento inferido a partir de las huellas observadas, mientras que un modelo de difusión emplea esta información para reconstruir una posible representación de la escena en el instante $t - \Delta$.



Fuente: Elaboración propia.

En este capítulo se describe la metodología utilizada para inferir eventos pasados a partir de evidencia visual residual. La propuesta sigue el paradigma de *Time-Reversed Imaging* y se adapta a un escenario multimodal basado en información RGB y térmica (TH).

El método busca reconstruir lo ocurrido en una escena instantes antes de la captura utilizando observaciones del presente. Para ello, el proceso se organiza en tres etapas: (i) adquisición multimodal, (ii) generación de descripciones semánticas y (iii) reconstrucción de la escena mediante modelos generativos guiados por lenguaje. La Figura 2 muestra el flujo general del sistema.

4.1. FORMULACIÓN DEL PROBLEMA

Sea $x_t^{RGB} \in \mathbb{R}^{H\times W\times 3}$ la imagen en el espectro visible y $x_t^{TH} \in \mathbb{R}^{H\times W}$ la imagen térmica, ambas capturadas de manera sincronizada en el instante actual t, después de una interacción entre un humano y el entorno. Sea además $\Delta > 0$ el desplazamiento temporal hacia el pasado que se desea inferir.

El problema considera dos objetivos principales: inferir el evento ocurrido en el instante $t-\Delta$, denotado como $y_{t-\Delta}$, y reconstruir una imagen RGB plausible $x_{t-\Delta}^{RGB}$ correspondiente al estado previo de la escena.
La relación probabilística entre estas variables se define como:

$$p(x_{t-\Delta}^{RGB}, y_{t-\Delta} \mid \mathcal{E}_t, \Delta), \tag{1}$$

donde la evidencia multimodal está dada por $\mathcal{E}_t = \{x_t^{RGB}, x_t^{TH}\}$. Debido a la complejidad de estimar esta distribución de manera directa, la formulación se divide en dos términos utilizando la regla de la cadena:

$$p(x_{t-\Delta}^{RGB}, y_{t-\Delta} \mid \mathcal{E}_t, \Delta) = p(x_{t-\Delta}^{RGB} \mid \mathcal{E}_t, y_{t-\Delta}, \Delta) \cdot p(y_{t-\Delta} \mid \mathcal{E}_t, \Delta). \tag{2}$$

Esta descomposición permite implementar un proceso en dos etapas: primero, la inferencia del evento pasado y, posteriormente, la reconstrucción de la escena condicionada a dicho evento.

4.2. DESCRIPCIÓN ESTRUCTURADA DEL EVENTO PASADO (SPED)

Para interactuar con los modelos de visión y lenguaje (VLM) utilizados en este trabajo, las instrucciones semánticas (*prompts*) se formularon en inglés. Esta decisión se tomó debido a que los modelos empleados fueron entrenados principalmente en dicho idioma y suelen mostrar un mejor desempeño en tareas de comprensión y generación de texto. Como resultado, las descripciones generadas tienden a ser más precisas y consistentes al representar el evento inferido[28].
Con el fin de reducir la ambigüedad inherente al problema, se propone el uso de una representación semántica estructurada denominada *Structured Past-Event Description (SPED)*. Esta representación es generada mediante un modelo visión-lenguaje (VLM), el cual recibe como entrada las observaciones actuales en RGB y térmico, y produce una descripción textual del evento más probable ocurrido en el pasado.

---

[28] Pritika ROHERA *et al.* «Better To Ask in English? Evaluating Factual Accuracy of Multilingual LLMs in English and Low-Resource Languages». En: *arXiv preprint arXiv:2504.20022* (2025).

Formalmente, se define como:

$$\hat{y}_{t-\Delta} = f_{\text{VLM}}(x_t^{RGB}, x_t^{TH}, \Delta, \text{SPED}), \tag{3}$$

donde $\hat{y}_{t-\Delta}$ corresponde a la hipótesis semántica estructurada. A diferencia de descripciones libres, SPED restringe el formato de salida mediante una plantilla fija, como se ilustra en la Figura 3.

Figura 3. Estructura del prompt SPED (Structured Past-Event Description).

SPED: Structured Past-Event Description

Using the provided thermal traces, infer the event that most likely occurred $\langle\Delta\rangle$ seconds earlier. Strictly follow this structure: The most recent event is that $\langle object\rangle$ was $\langle action\rangle$ [optionally involving $\langle object_2\rangle$ or $\langle action_2\rangle$].

Fuente: Elaboración propia.

El uso de información térmica resulta fundamental en esta formulación, ya que permite identificar patrones de transferencia de calor asociados a interacciones recientes, proporcionando evidencia física que complementa la información visual.

4.3. ARQUITECTURA DE AJUSTE FINO PARA EL MÓDULO DESCRIPTOR

Dado que el entrenamiento de un modelo generativo visual de extremo a extremo resulta computacionalmente inviable por la ambigüedad de las huellas térmicas[29], la reconstrucción visual se delega a modelos de difusión de gran escala ya consolidados[30]. Por consiguiente, el esfuerzo arquitectónico propuesto se concentra exclusivamente en optimizar el módulo descriptor.

Con el fin de evaluar si era posible generar el SPED utilizando modelos preentrenados de

[29] TANG *et al.* Ver n. 7; BERTERO y BOCCACCI, ver n. 10.

[30] BROOKS, HOLYNSKI y EFROS, ver n. 26; ZHANG, RAO y AGRAWALA, ver n. 21.

menor tamaño en entornos de cómputo local, se estudiaron dos estrategias:

1. **Inicialización textual:** un proceso de refinamiento donde un modelo de lenguaje recibe descripciones preliminares y las convierte al formato estructurado SPED.

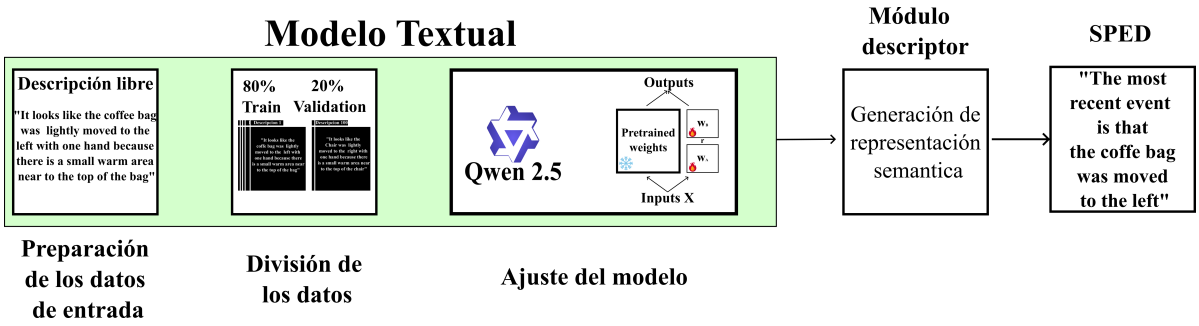
2. **Información multimodal:** un esquema de inferencia directa en el que un modelo visión-lenguaje procesa simultáneamente las imágenes RGB, las imágenes térmicas (TH) y el intervalo temporal ($\triangle$) para generar la hipótesis SPED.

4.3.1. Inicialización textual: Refinamiento semántico La inicialización textual se utilizó para convertir descripciones libres en sentencias con el formato SPED. En esta etapa, el modelo no recibe directamente la evidencia física de la escena; su función consiste en reorganizar y estructurar información previamente descrita para obtener una representación uniforme. Esta etapa facilita el uso posterior de las descripciones como entrada del modelo generativo de reconstrucción.

Para construir el conjunto de datos utilizado en esta etapa, fue necesario generar manualmente las descripciones de entrada. Para ello, dos observadores revisaron cada escena del conjunto de datos utilizando simultáneamente las imágenes RGB y térmicas. Con base en la información disponible en ambas modalidades, redactaron descripciones generales de lo que ocurría en cada escena.

Estas descripciones constituyeron la entrada del modelo de lenguaje, cuyo objetivo fue transformarlas al formato SPED empleado en el sistema.

Figura 4. **Esquema de inicialización textual.** Representación de la transformación de descripciones libres hacia el formato estructurado.



Fuente: Elaboración propia.

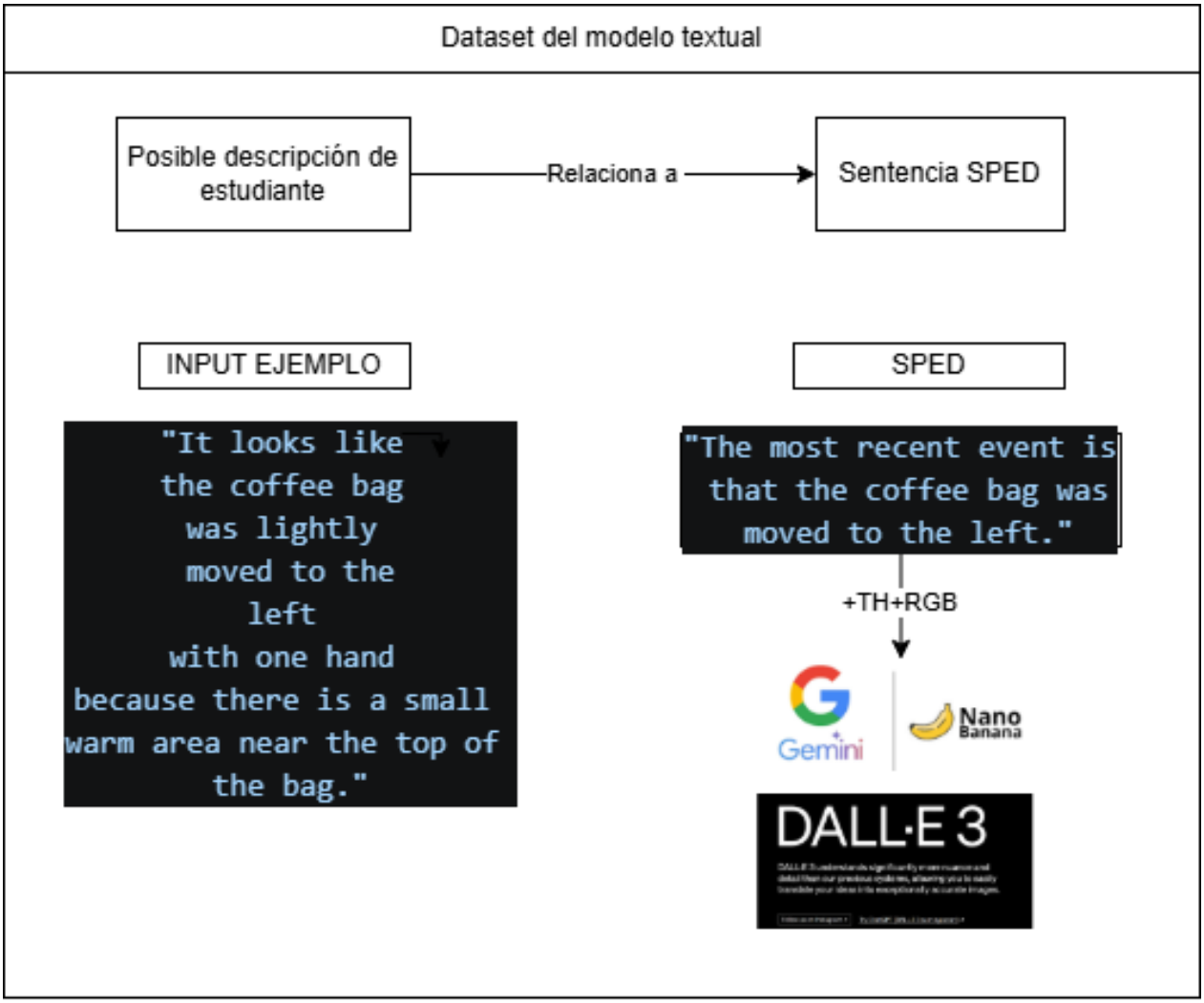
Para esta tarea se implementó `Qwen2.5-3B-Instruct`[31], seleccionado por su equilibrio entre eficiencia paramétrica y alta capacidad de seguimiento de instrucciones (*instruction-following*). El diseño del modelo y su flujo de procesamiento se ilustran en la Figura 4 y la Figura 5. A nivel de optimización, el modelo fue ajustado empleando la técnica de Adaptación de Bajo Rango (LoRA)[32]. Al intervenir únicamente una fracción reducida de los parámetros entrenables, se preservó el conocimiento lingüístico del modelo base, previniendo el sobreajuste frente al conjunto de datos local. Las especificaciones paramétricas completas y las gráficas de convergencia de este proceso se detallan en el Anexo B.

---

[31] An YANG *et al.* «Qwen2 Technical Report». En: *arXiv preprint arXiv:2407.10671* (2024).

[32] Edward J. HU *et al.* «LoRA: Low-Rank Adaptation of Large Language Models». En: *International Conference on Learning Representations*. 2022.

Figura 5. **Arquitectura del modelo textual.** Esquema de integración de la red `Qwen2.5-3B-Instruct` adaptada para la tarea de traducción de descripciones libres a representaciones SPED mediante instrucciones estandarizadas.

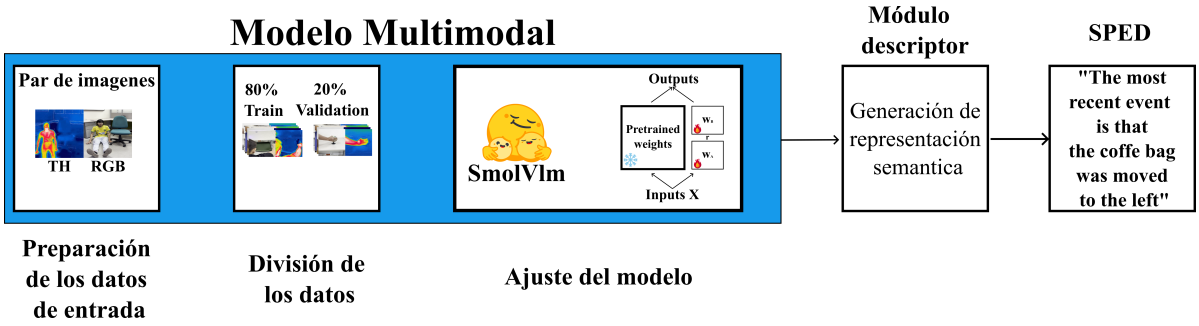


Fuente: Elaboración propia.

4.3.2. Información Multimodal: Asimilación directa de evidencia física A diferencia de la inicialización textual, el enfoque de información multimodal se diseñó para procesar directamente la evidencia empírica del problema. El sistema recibe la imagen RGB, el mapa de calor residual (TH) y el horizonte temporal (Δ) con el propósito de deducir autónomamente la representación SPED. En este esquema, la red neuronal debe aprender a correlacionar patrones geométricos visuales con distribuciones termodinámicas para

formular una hipótesis causal sobre el evento previo.

Figura 6. **Esquema de inicialización multimodal.** Representación de la convergencia de entrada RGB y térmica hacia la sentencia SPED.



Fuente: Elaboración propia.

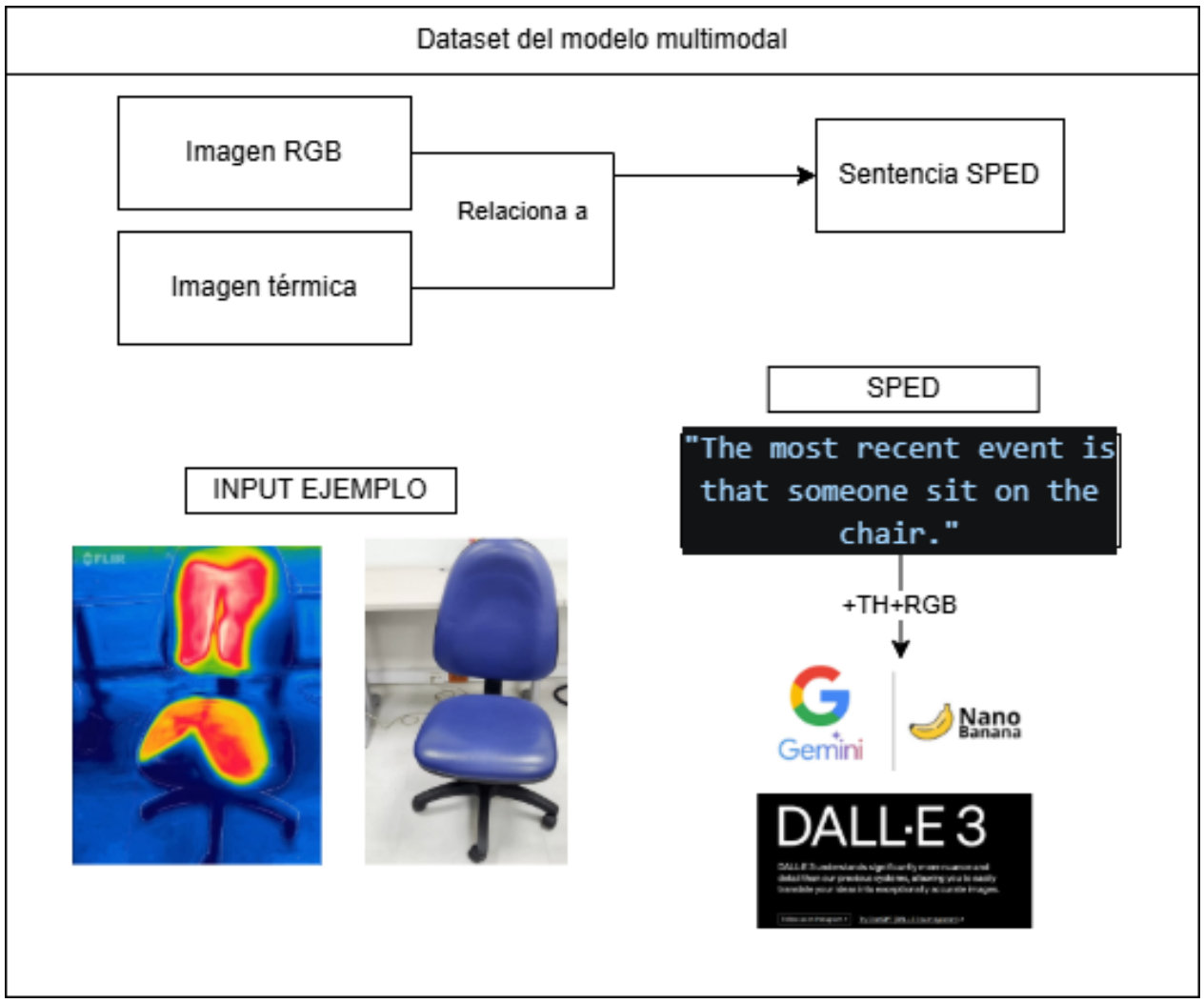
Para la implementación de esta ruta se adoptó la arquitectura `SmolVLM-Instruct`[33], respondiendo a un estricto criterio de eficiencia paramétrica que garantiza la viabilidad del procesamiento concurrente de tensores de alta dimensionalidad. La arquitectura del flujo de datos y la integración del modelo se muestran en las Figuras 6 y 7. De forma similar a los experimentos anteriores, el ajuste del modelo se realizó mediante LoRA[34], con el fin de preservar las representaciones visuales aprendidas durante el preentrenamiento. Los detalles del hardware utilizado, los hiperparámetros de entrenamiento y las curvas obtenidas durante el proceso se presentan en el Anexo B.

---

[33] HUGGING FACE. *SmolVLM-Instruct*. `https://huggingface.co/HuggingFaceTB/SmolVLM-Instruct`. [Model card en Hugging Face. Fecha de consulta: 5 de abril de 2026]. 2025.

[34] HU *et al.* Ver n. 32.

Figura 7. **Arquitectura del flujo multimodal.** Diagrama conceptual del procesamiento concurrente donde las modalidades RGB y térmica convergen para condicionar la inferencia de la representación SPED.



Fuente: Elaboración propia.

Esta etapa da respuesta al Objetivo Específico 2, ya que permite construir un modelo descriptor capaz de identificar los objetos involucrados en una interacción y generar una descripción semántica del evento previo.

4.4. RECONSTRUCCIÓN DE LA ESCENA PASADA MEDIANTE DIFUSIÓN GUIADA

Una vez obtenida la descripción semántica $\hat{y}_{t-\Delta}$, el siguiente paso consiste en generar una representación visual de la escena correspondiente al instante pasado. Este proceso puede expresarse mediante:

$$p(x_{t-\Delta}^{RGB} \mid \mathcal{E}_t, \hat{y}_{t-\Delta}), \tag{4}$$

que se modela utilizando un sistema de difusión condicionado. La generación comienza con un vector latente aleatorio $z_T \sim \mathcal{N}(0, I)$ y avanza mediante sucesivas etapas de eliminación de ruido:

$$z_{t-1} = g_\phi(z_t, \mathcal{E}_t, \hat{y}_{t-\Delta}), \tag{5}$$

hasta obtener una representación latente final que posteriormente se decodifica como $\hat{x}_{t-\Delta}^{RGB} = D(z_0, \hat{y}_{t-\Delta})$. La información semántica para el proceso se incorpora mediante el *prompt* mostrado en la Figura 8.

Figura 8. Estructura del prompt de reconstrucción para la edición de la imagen RGB.

Reconstruction Prompt

Edit the RGB image to reconstruct the scene as it appeared $<\Delta>$ seconds earlier, preserving the environment, lighting, viewpoint, and color distribution. Assume that the recent event was that $<\hat{y}_{t-\Delta}>$

Fuente: Elaboración propia.

La reconstrucción combina la evidencia térmica disponible con la descripción semántica inferida, generando una estimación visual del estado previo de la escena. Esta etapa da respuesta al Objetivo Específico 3, ya que utiliza la descripción generada por el modelo descriptor para guiar el proceso de reconstrucción mediante técnicas de difusión.

5. SIMULACIONES Y RESULTADOS

Para evaluar el marco de inferencia temporal inversa propuesto, se definió un conjunto de experimentos basado en métricas complementarias.

En total, se realizaron tres experimentos de ablación para analizar diferentes componentes del método. Estos experimentos consideran variaciones en el modelo generativo utilizado y estudian el efecto de distintos intervalos temporales (Δt), prestando especial atención a ventanas cortas de reconstrucción de hasta dos minutos.

Todas las evaluaciones se llevaron a cabo sobre el conjunto de datos TRACE-HEI. El método se evaluó bajo un esquema libre de entrenamiento, por lo que el conjunto de datos se empleó únicamente para pruebas, sin realizar procesos de aprendizaje ni ajuste de parámetros. Los resultados reportados corresponden al promedio de cinco ejecuciones independientes.

5.1. CONJUNTO DE DATOS

Se construye un conjunto de datos multimodal compuesto por secuencias sincronizadas en los espectros visible (RGB) y térmico (TH), capturadas durante y después de interacciones controladas entre humanos y el entorno. Para la adquisición de los datos se utilizan dos sensores: (i) una cámara RGB de alta resolución (iPhone 15, 4032×3024, rango visible 400–700 nm) y (ii) una cámara térmica FLIR ONE Pro (160×120, infrarrojo de onda larga 8–14 μm). La Figura 10**(a)** presenta la configuración utilizada para la captura de datos y la posición de los sensores.

Luego de la captura, las secuencias son sincronizadas y alineadas geométricamente para reducir diferencias espaciales entre ambas modalidades. También se aplica una normalización sobre las imágenes RGB y térmicas con el fin de mejorar la correspondencia entre ellas. Además, cada secuencia se normaliza para reducir las variaciones asociadas a cambios de iluminación y diferencias en el rango dinámico de los sensores. Una descripción detallada de la configuración de adquisición y del pipeline de procesamiento se presenta en el Anexo A.

El protocolo de adquisición se divide en dos fases:

Figura 9. **Ejemplos visuales del conjunto de datos TRACE-HEI.** Muestras representativas del sistema de adquisición multimodal que ilustran diversas interacciones humano–entorno, capturadas en las modalidades RGB y térmica.

Fuente: Elaboración propia.

- **Fase de interacción (0–30 s):** Se realiza una interacción controlada entre el sujeto y el entorno, mantenida durante un máximo de 30 segundos. Esta fase define el

evento pasado de referencia y el estado de la escena que se desea reconstruir.

- **Fase de decaimiento (30–120 s):** Una vez finalizada la interacción, se registra una secuencia multimodal durante 120 segundos con el objetivo de analizar la evolución temporal y disipación de las huellas físicas. Esta fase permite estudiar la persistencia de las señales térmicas y determinar hasta qué retraso temporal Δ es posible realizar inferencias fiables sobre el estado pasado de la escena.

El conjunto de datos está compuesto por 100 escenas multimodales que abarcan cuatro categorías principales de interacción: transiciones de sentarse/levantarse, interacciones de contacto, manipulación de objetos e interacciones con generación de residuos. Como se observa en la Figura 10**(b)**, cada categoría contiene 13 escenas, manteniendo una distribución equilibrada entre las acciones evaluadas.

Los objetos utilizados buscan representar situaciones comunes en entornos interiores. Entre ellos se incluyen sillas, teclados, mesas, bolsas, recipientes y superficies de suelo. La variedad de objetos introduce cambios en la geometría, el área de contacto y el comportamiento térmico de cada interacción, afectando tanto la formación como la disipación de las huellas térmicas.

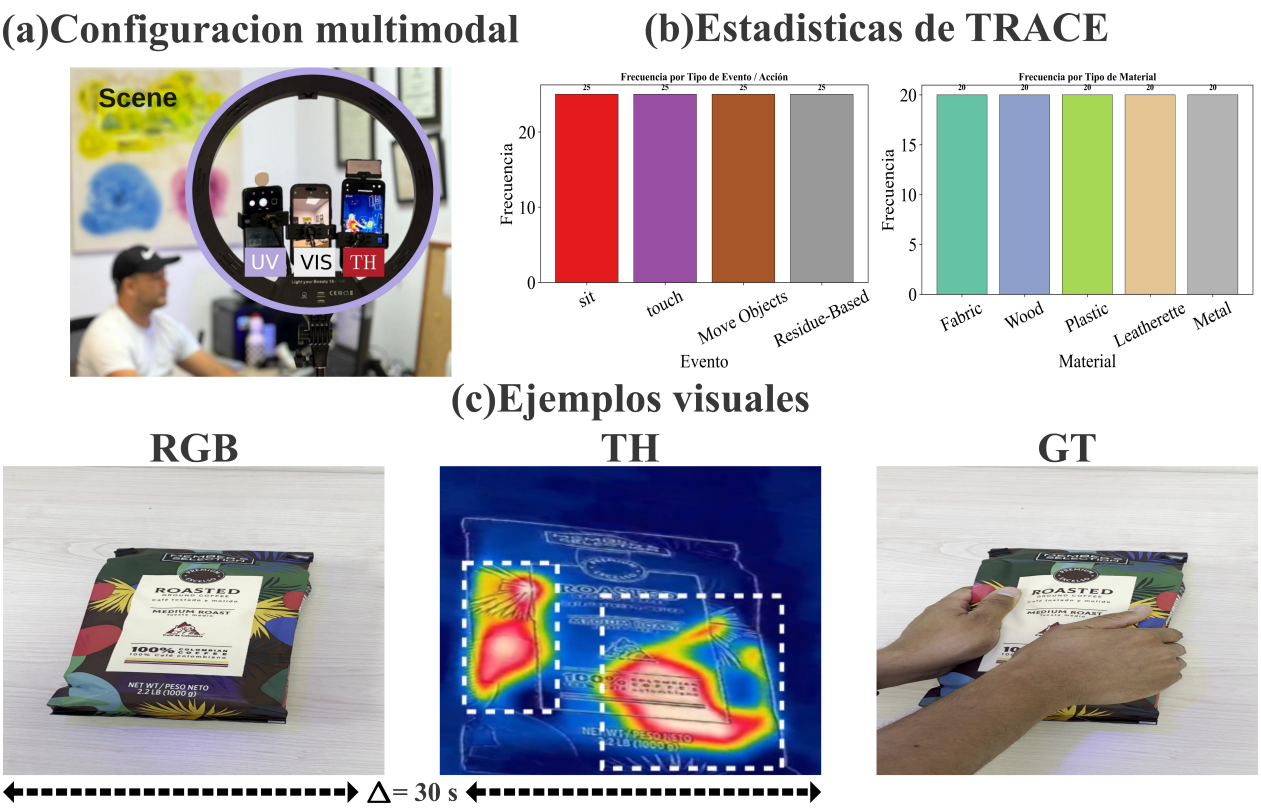
También se consideraron materiales distintos, como plástico, metal, tela, madera y cerámica. Debido a sus propiedades térmicas, cada material conserva y disipa el calor de manera diferente, produciendo patrones térmicos con comportamientos variados.

En este contexto, la información térmica permite observar cómo el calor se transfiere y desaparece después del contacto físico. Estas variaciones contienen información útil para inferir interacciones recientes.

La Figura 10**(c)** muestra un ejemplo representativo del conjunto de datos. En este caso, la imagen fue capturada $\Delta = 30$ segundos después de la interacción, cuando las huellas térmicas todavía pueden observarse con claridad. Se incluye además la imagen de referencia (GT), que representa el estado de la escena durante la interacción. Este ejemplo corresponde a una interacción de contacto con un objeto de material plástico.

Finalmente, esta subsección da cumplimiento al Objetivo Específico 1 del trabajo de grado, al construir y caracterizar un conjunto de datos multimodal compuesto por pares de imágenes RGB y térmicas alineadas, capturadas en escenarios controlados de interacción (sentarse, tocar y apoyarse), e incluyendo distintos intervalos de degradación térmica

Figura 10. **Resumen del conjunto de datos.** (a) Configuración de adquisición multimodal con sensores RGB y térmico sincronizados para capturar huellas residuales posteriores a la interacción. (b) Estadísticas del conjunto de 52 escenas (acciones, objetos y materiales). (c) Ejemplo representativo a un retraso temporal de $\Delta = 30\,\text{s}$, donde la modalidad térmica captura información residual del evento.



Fuente: Elaboración propia.

(0, 5, 15, 30 y 120 segundos), lo cual establece una base sólida para los procesos de entrenamiento, validación y evaluación del sistema propuesto.

5.2. MÉTRICAS

La evaluación de la reconstrucción temporal inversa se realiza mediante tres niveles complementarios: fidelidad a nivel de píxel, consistencia perceptual basada en características y corrección semántica a alto nivel. Esta organización permite analizar distintos aspectos de las reconstrucciones generadas, ya que la calidad del resultado no depende únicamente de la similitud entre píxeles, sino también de la coherencia estructural y semántica

de la escena reconstruida.

5.2.1. Métricas de bajo nivel. Se emplean las métricas Peak Signal-to-Noise Ratio (PSNR)[35] y Structural Similarity Index (SSIM)[36] para cuantificar la fidelidad de reconstrucción en el espacio de píxeles.

PSNR mide las discrepancias absolutas de intensidad entre la imagen reconstruida $\hat{x}_{t-\Delta}^{RGB}$ y la imagen de referencia del pasado $x_{t-\Delta}^{RGB}$, proporcionando una evaluación estándar a nivel de señal. Por su parte, SSIM evalúa la similitud estructural considerando luminancia, contraste y estructura local, lo cual permite capturar degradaciones perceptuales que no son evidentes únicamente con métricas basadas en error absoluto.

5.2.2. Métricas a nivel de características. Para evaluar la similitud perceptual más allá del espacio de píxeles, se utiliza la métrica Learned Perceptual Image Patch Similarity (LPIPS)[37], la cual calcula distancias en espacios de características profundas y presenta una mayor correlación con la percepción humana. Esta métrica resulta especialmente relevante en este contexto, ya que la reconstrucción temporal puede generar resultados semánticamente correctos pero con desalineaciones a nivel de píxel.

Adicionalmente, se calcula una medida de similitud basada en CLIP[38] entre la imagen generada y la descripción estructurada del evento pasado. Esta métrica permite evaluar la alineación semántica entre la hipótesis textual inferida $\hat{y}_{t-\Delta}$ y la reconstrucción visual, garantizando coherencia entre el razonamiento guiado por lenguaje y la síntesis de la imagen.

---

[35] Zhou WANG *et al.* «Image quality assessment: From error visibility to structural similarity». En: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600-612.

[36] WANG *et al.* Ver n. 35.

[37] Richard ZHANG *et al.* «The unreasonable effectiveness of deep features as a perceptual metric». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 586-595.

[38] RADFORD *et al.* Ver n. 8.

5.2.3. Métricas de alto nivel. Para medir la concordancia semántica global, se reporta la métrica Overall Accuracy (OA)[39]. En particular, se utiliza el modelo Segment Anything Model (SAM)[40] para obtener máscaras de objetos tanto en la imagen reconstruida como en la escena de referencia. La métrica OA se calcula como el acuerdo a nivel de píxel entre las regiones segmentadas predichas y las de referencia, permitiendo evaluar si los objetos son correctamente identificados y ubicados dentro de la escena.

Es importante destacar que esta métrica prioriza la consistencia estructural y relacional sobre el refinamiento preciso de bordes, enfocándose en la correspondencia global de los objetos.

Para complementar esta evaluación, se emplea un detector preentrenado basado en YOLOv11[41] con el fin de localizar los principales objetos involucrados en la interacción. Posteriormente, se calcula la métrica Intersection-over-Union (IoU)[42] entre las cajas delimitadoras predichas y las de referencia. Mientras que OA mide la consistencia semántica global, IoU evalúa específicamente la precisión en la localización espacial de los objetos clave dentro de la escena.

5.3. EXPERIMENTOS

5.3.1. Evaluación exploratoria de los módulos descriptores (Ajuste Fino) Con base en la arquitectura diseñada para la generación del formato SPED, detallada en el marco metodológico, se procedió a evaluar la capacidad empírica de los modelos ajustados localmente (Aproximación Textual y Aproximación Multimodal). Para la ejecución de este ajuste fino, se empleó un subconjunto de entrenamiento compuesto por 100 escenas, lo que equivale a un volumen aproximado de 1800 imágenes multimodales extraídas de

---

[39] Abdel Aziz TAHA y Allan HANBURY. «Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool». En: *BMC Medical Imaging* 15.1 (2015), pág. 29.

[40] Alexander KIRILLOV, Eric MINTUN, Nikhila RAVI *et al.* «Segment Anything». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 4015-4026.

[41] Glenn JOCHER *et al. Ultralytics YOLOv11*. `https://github.com/ultralytics/ultralytics`. 2024.

[42] Mark EVERINGHAM *et al.* «The Pascal Visual Object Classes (VOC) challenge». En: *International Journal of Computer Vision* 88.2 (2010), pp. 303-338.

interacciones físicas controladas.

El objetivo de esta evaluación fue analizar si los modelos de escala intermedia, entrenados bajo las restricciones de datos disponibles, eran capaces de generar descripciones suficientemente precisas para apoyar la etapa posterior de reconstrucción visual.

5.3.1.1. Análisis cuantitativo y comparativa SOTA Los experimentos de ajuste fino realizados sobre el conjunto de 100 escenas mostraron que es posible adaptar modelos de menor tamaño para generar descripciones compatibles con el formato SPED. Sin embargo, cuando estos modelos se incorporaron al flujo completo de reconstrucción, su desempeño fue limitado, especialmente en tareas que requerían una interpretación más precisa de la información física disponible.

El Cuadro 1 compara el desempeño de las arquitecturas ajustadas localmente con modelos ampliamente utilizados en la literatura reciente, incluyendo GPT-5[43], Gemini 3 Pro[44], Claude 3.5[45], Qwen3.5 Plus[46] y LLaVA-NeXT[47]. Los resultados demuestran una discrepancia significativa: mientras los modelos de gran escala alcanzan precisiones superiores al 85 % en la identificación de acciones y objetos, los modelos locales ajustados promedian rendimientos sustancialmente menores en tareas complejas de razonamiento exacto.

---

[43] OPENAI. *GPT-5 Technical Report*. `https://openai.com/research/gpt-5`. [Fecha de consulta: 10 de abril de 2026]. 2025.

[44] GOOGLE DEEPMIND. *Gemini 3: Next-Generation Multimodal Models*. `https://deepmind.google/technologies/gemini/`. [Fecha de consulta: 10 de abril de 2026]. 2026.

[45] ANTHROPIC. *Claude 3.5: Next-generation AI assistant*. `https://www.anthropic.com/news/claude-3-5-sonnet`. [Fecha de consulta: 28 de marzo de 2026]. 2024.

[46] QWEN TEAM *et al.* «Qwen3.5: Advancing Large Vision-Language Models». En: *arXiv preprint* (2025).

[47] Haotian LIU *et al. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge*. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`. [Fecha de consulta: 10 de abril de 2026]. 2024.

Cuadro 1. **Evaluación de inferencia semántica en modelos visión-lenguaje.** Desempeño comparativo de siete arquitecturas VLM al generar descripciones estructuradas de eventos pasados a partir de observaciones multimodales ($\Delta = 30$s). Se indica entre paréntesis la escala paramétrica de cada modelo (marcando como No revelado aquellos de arquitectura cerrada). Los resultados se reportan como media $\pm$ desviación estándar sobre tres ejecuciones independientes.

Modelo VLM	**Acción ↑**	**Objeto ↑**	**Material ↑**	**Coincidencia Exacta ↑**	**Top-3 ↑**
GPT-5 (No revelado)	**88.4 ± 0.5**	91.2 ± 0.3	**85.7 ± 0.8**	**78.3 ± 0.9**	**94.6 ± 0.2**
Gemini 3 Pro (No revelado)	86.9 ± 0.6	**91.5 ± 0.4**	84.2 ± 0.9	76.1 ± 1.1	93.8 ± 0.3
Claude 3.5 (No revelado)	84.2 ± 0.7	87.8 ± 0.5	82.1 ± 1.0	72.5 ± 1.2	91.2 ± 0.4
Qwen3.5 Plus (No revelado)	79.5 ± 1.2	82.3 ± 1.1	76.4 ± 1.5	65.8 ± 1.8	86.4 ± 0.9
LLaVA-NeXT (34B)	72.1 ± 1.5	75.4 ± 1.4	68.9 ± 1.9	58.2 ± 2.1	79.1 ± 1.2
Qwen2.5 Instruct - Textual local (3B)	83.6 ± 1.4	74.8 ± 1.2	22.5 ± 1.1	0 ± 0	41.7 ± 1.0
SmolVLM Instruct - Multimodal local (500M)	52.3 ± 1.6	38.7 ± 1.3	12.4 ± 0.9	0 ± 0	49.8 ± 0.8

Fuente: Elaboración propia a partir de las pruebas de inferencia ejecutadas.

Para comprender a profundidad la naturaleza de esta brecha de rendimiento y las limitaciones inferenciales en los modelos locales, a continuación se detallan los resultados cualitativos obtenidos para cada aproximación experimental.

5.3.2. Comparación semántica entre el modelo textual y el modelo multimodal Con el fin de contrastar el desempeño de ambos enfoques, se realizó una evaluación semántica conjunta tomando como referencia el mismo *ground truth* y aplicando un criterio de adjudicación manual. Este criterio privilegió la coincidencia del núcleo del evento y del objeto principal, incluso en los casos donde la formulación lingüística no coincidía de manera literal con la oración ideal. Al revisar las escenas evaluadas, se encontraron diferencias claras entre ambos modelos. El modelo textual fue más consistente al momento de describir la acción principal y el objeto involucrado. Por otro lado, el modelo multimodal generó respuestas más inestables y, en algunos casos, enfocó la descripción en elementos que no correspondían al evento principal.

En escenas con objetos manipulados manualmente, como bolsas o pocillos, ambos modelos lograron recuperar parte de la información presente en la interacción. Sin embargo, las descripciones generadas por el modelo textual tendieron a ser más directas y cerca-

nas al evento observado. En contraste, el modelo multimodal mostró una mayor variabilidad en la descripción de las acciones y con frecuencia omitió información adicional presente en la escena. Esta diferencia fue más visible en las interacciones asociadas al uso de sillas, donde el modelo textual identificó correctamente acciones como *sat on*, mientras que el modelo multimodal las describió en ocasiones como movimientos o cambios de posición.

El Cuadro 2 presenta algunos ejemplos que permiten comparar el comportamiento de ambas aproximaciones frente a la misma evidencia física.

Cuadro 2. **Comparación semántica por escena entre el modelo textual y multimodal.** Selección de casos representativos que ilustran la capacidad de cada modelo para mantener el núcleo del evento frente a la evidencia física.

Escena	***Ground truth***	**Modelo textual**	**Modelo multimodal**	**Lectura comparativa**
16	Metal cup moved	Cup moved to the right with a hand	Mug on the table grasped and moved slightly	Ambos recuperan la acción; el textual es más directo y el multimodal añade complementos locativos.
20	Metal cup touched	Red metal cup grabbed from above	All four cups on the table were used	El textual recupera acción, objeto y material; el multimodal se vuelve difuso y sobregeneraliza.
24	Fabric chair sat on	Person sat on the chair	Chair was moved slightly to the right	El textual coincide con el evento central; el multimodal desvía la interpretación hacia movimiento físico.
26	Fabric chair sat on	A person sat in the chair	Two people moved from the chair to the wall	El textual mantiene el núcleo semántico; el multimodal desplaza la lectura hacia una transición espacial.

Fuente: Elaboración propia.

En conjunto, los resultados mostraron un mejor desempeño del modelo textual en la identificación de la acción principal y del objeto involucrado en cada escena. Aunque el modelo multimodal logró reconocer algunos elementos correctamente, presentó errores con mayor frecuencia al describir la acción realizada y omitió información adicional en varios casos. Como resultado, las descripciones generadas por el modelo textual mantuvieron una mayor correspondencia con el *ground truth*, mientras que las salidas del modelo multimodal fueron menos consistentes entre escenas.

5.3.2.1. Análisis cualitativo: Aproximación textual Para este análisis se revisaron las predicciones generadas por el modelo textual en las escenas 16 a 26 y se compararon con el *ground truth* utilizado en la evaluación. El análisis no se centró únicamente en la coincidencia literal entre las descripciones, sino también en la conservación del significado de la escena, considerando la acción principal, el objeto involucrado y atributos adicionales como el material o la persona. Este criterio resulta relevante porque una descripción puede representar correctamente un evento aunque utilice una redacción diferente a la referencia.

En términos generales, el modelo logró identificar adecuadamente los elementos principales de las interacciones. La Escena 16, mostrada en el Cuadro 3, evidencia que el modelo pudo identificar tanto la acción principal (*moved to the right*) como el objeto asociado a la interacción. Aunque el material del recipiente no fue mencionado, la descripción siguió siendo coherente con el evento observado.

En la Escena 19 se encontró la diferencia más importante. En este caso, el modelo interpretó la acción como *grabbed* en vez de un desplazamiento. Sin embargo, el objeto detectado sí coincidió con el contenido de la escena.

El modelo también mostró un comportamiento estable en interacciones relacionadas con el uso de mobiliario. La Escena 25 constituye un ejemplo de ello, ya que la descripción mantuvo el concepto principal de la interacción (*sat in the chair*) e identificó correctamente al sujeto de la acción (*the girl*). Estos resultados sugieren que el modelo prioriza los elementos más relevantes de la escena, manteniendo una buena correspondencia con el *ground truth* incluso cuando algunos atributos secundarios no son incluidos en la descripción.

Cuadro 3. **Ejemplos representativos del análisis cualitativo del modelo textual.**

Escena	*Ground truth*	Salida del modelo textual	Lectura
16	The most recent event is that the metal cup was moved.	The most recent event is that the cup was moved to the right with a hand.	Recupera correctamente la acción y el objeto.
19	The most recent event is that the metal cup was moved.	The most recent event is that the small cup was grabbed from the top left.	Mantiene el objeto, pero desvía la acción principal.
25	The most recent event is that the fabric chair was sat on.	The most recent event is that the girl sat in the chair.	Alta coincidencia semántica; recupera exitosamente a la persona.

Fuente: Elaboración propia.

En conclusión, la aproximación textual mostró un comportamiento consistente al traducir observaciones visuales en proposiciones SPED. Aunque se observaron omisiones en algunos atributos secundarios y ciertas diferencias en la descripción de acciones específicas, el modelo logró conservar la información principal de la mayoría de las escenas, identificando correctamente el evento central y el objeto involucrado.

5.3.2.2. Análisis cualitativo: Aproximación multimodal La evaluación del modelo multimodal se realizó sobre el mismo conjunto de escenas (16 a 26), utilizando el mismo criterio de validación semántica. En esta etapa, la revisión manual resultó necesaria debido a que algunas métricas automáticas penalizaban descripciones que conservaban el significado del evento, pero diferían de la referencia en su estructura gramatical. En varios casos, el modelo mencionó correctamente el objeto principal dentro de respuestas más extensas, lo que produjo errores durante la evaluación automática.

Al revisar manualmente los resultados, se encontró que el modelo multimodal logró describir varias interacciones físicas simples. En el Cuadro 4, la Escena 16 muestra un ejemplo de este comportamiento. El modelo identificó la acción (*grasped and moved slightly*) y el

objeto asociado, aunque agregó detalles adicionales que hicieron la descripción menos precisa.

También aparecieron errores importantes. En la Escena 20, el modelo interpretó que todos los recipientes de la escena habían sido utilizados (*all four cups... were used*), cuando la huella térmica correspondía únicamente a uno de ellos. En este caso, la descripción no coincidió con la interacción observada.

Las mayores dificultades aparecieron en escenas donde era necesario interpretar el uso del objeto. Por ejemplo, en la Escena 25 el modelo no relacionó la huella térmica con el uso de la silla y generó una descripción enfocada en elementos ajenos al evento principal (*the woman held the phone*). Esto muestra limitaciones para asociar correctamente la evidencia térmica con la interacción que produjo dicha huella.

Cuadro 4. **Ejemplos representativos del análisis cualitativo del modelo multimodal.**

Escena	***Ground truth***	**Salida del modelo multimodal**	**Lectura**
16	The most recent event is that the metal cup was moved.	The most recent event is that the mug on the table was grasped and moved slightly with one hand.	Semánticamente aceptable, aunque con estructura dispersa.
20	The most recent event is that the metal cup was touched.	The most recent event is that all four cups on the table were used.	Sobregeneraliza la interacción y pierde precisión focal.
25	The most recent event is that the fabric chair was sat on.	The most recent event is that the woman held the phone.	Desvía la interpretación causal e introduce alucinaciones.

Fuente: Elaboración propia.

A partir de las escenas evaluadas, se observa que el modelo multimodal local fue menos estable que el modelo textual. Aunque describió correctamente algunas interacciones simples, presentó problemas cuando debía interpretar el contexto de uso de los objetos o concentrarse en una región térmica específica. En comparación, el modelo textual mantuvo una mayor estabilidad en la identificación del evento principal y de los elementos más relevantes de cada escena.

5.3.2.3. Consideraciones finales del módulo descriptor Las diferencias observadas entre los modelos parecen estar relacionadas con su escala y capacidad de representación. Durante los experimentos locales, las limitaciones del hardware disponible hicieron necesario utilizar arquitecturas ligeras, como SmolVLM con aproximadamente 500 millones de parámetros. Aunque estos modelos presentan ventajas en términos de eficiencia computacional, mostraron dificultades para interpretar correctamente patrones asociados a la disipación térmica y ciertas relaciones espaciales presentes en el problema de reconstrucción.

En este tipo de reconstrucción, el modelo generador de imágenes depende estrictamente de la descripción semántica que recibe. Si el módulo descriptor comete un error o emite una instrucción incompleta, como ocurrió con los modelos locales debido a su limitada capacidad de ajuste, esta falla inicial se amplifica drásticamente. El resultado es que el sistema termina generando alucinaciones visuales o alterando por completo la estructura física de la escena reconstruida.

Por este motivo, y respaldados por la superioridad evidenciada en los resultados cuantitativos, se concluye que la tarea exige una altísima capacidad de comprensión del entorno que solo las arquitecturas de escala industrial pueden ofrecer. En consecuencia, se decidió integrar GPT-5 como el motor principal del módulo descriptor para el flujo de producción del proyecto. Esta decisión metodológica garantiza la precisión semántica requerida para guiar correctamente la síntesis visual y asegurar la coherencia física de los resultados finales.

Cuadro 5. **Estudio de ablación de modalidades de entrada para reconstrucción visual ($\Delta = 30$s).** Análisis de desempeño al integrar progresivamente la señal térmica (TH) y los descriptores semánticos (Libre y SPED). El mejor desempeño para cada métrica se reporta en **negrita**, mientras que el segundo mejor resultado se subraya.

RGB	TH	Descript.		Bajo Nivel		Nivel de Características		Alto Nivel	
		Libre	SPED	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP ↑	OA ↑	IoU ↑
✓	✗	✗	✗	11.135 ± 2.007	0.477 ± 0.190	0.446 ± 0.108	62.264 ± 8.891	0.617 ± 0.301	0.413 ± 0.299
✓	✓	✗	✗	14.490 ± 5.003	0.558 ± 0.234	0.537 ± 0.157	77.550 ± 2.898	0.818 ± 0.167	0.567 ± 0.251
✓	✓	✓	✗	17.056 ± 3.372	0.682 ± 0.156	0.320 ± 0.184	**93.938 ± 2.015**	0.765 ± 0.232	0.624 ± 0.193
✓	✓	✗	✓	**18.268 ± 2.733**	**0.718 ± 0.146**	**0.297 ± 0.195**	92.755 ± 3.177	**0.854 ± 0.249**	**0.739 ± 0.177**

Fuente: Elaboración propia a partir de los resultados del estudio de ablación.

5.3.3. Estudio de ablación de entrada para reconstrucción Con el objetivo de analizar la contribución de cada componente del marco propuesto, se realiza un estudio de ablación sistemático enfocado en dos factores principales: (i) el impacto de la modalidad de sensado térmico en el generador y en el descriptor semántico, y (ii) el papel de las descripciones estructuradas de eventos pasados en la restricción del proceso generativo. Para garantizar una comparación controlada, se fija el horizonte de reconstrucción en $\Delta = 30$ segundos, es decir, la tarea consiste en reconstruir el estado de la escena RGB 30 segundos antes de la observación multimodal actual. Para la inferencia de la descripción del evento pasado, se emplea ChatGPT-5.2[48] como modelo visión-lenguaje (VLM). Para la generación de imágenes, se utiliza Gemini 2.5[49] como modelo generativo guiado por dichas descripciones. La evaluación de distintos modelos generativos se presenta en la siguiente subsección.

El cuadro 5 presenta los resultados del estudio de ablación. En las primeras configuraciones (filas 1–2), no se emplea ninguna descripción explícita del evento pasado. En estos casos, el *prompt* de reconstrucción incluye únicamente su componente base (descrito en la Sección correspondiente), incorporando la instrucción: ‘‘`taking into account the attached <RGB/Thermal>images.`’’

Cuando se incorporan descripciones semánticas (filas 5–6), se evalúan dos estrategias. La primera corresponde a una descripción en formato libre (*free-form*), donde se utiliza únicamente la parte inicial del prompt SPED sin imponer estructura. La segunda estrategia corresponde al enfoque propuesto SPED, en el cual se utiliza una plantilla estructurada completa que fuerza una descomposición explícita del evento pasado, reduciendo la ambigüedad en la interpretación.

Los resultados obtenidos muestran que el uso exclusivo de imágenes RGB produce los valores más bajos en todas las métricas evaluadas. Esto era esperable, ya que la imagen visible no contiene información directa sobre eventos ocurridos antes de la captura. En la

---

[48] OPENAI. *DALL·E 3*. `https://openai.com/dall-e-3`. [Fecha de consulta: 30 de septiembre de 2025]. 2023.

[49] GOOGLE. *Introducing Gemini 2.5 Flash Image (aka Nano Banana)*. `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/`. [Fecha de consulta: 30 de septiembre de 2025]. 2025.

mayoría de los casos, el modelo simplemente conserva el estado actual de la escena.

Al incorporar información térmica, el desempeño mejora en todas las métricas. Se observaron incrementos de hasta 2 dB en PSNR, mejoras en métricas perceptuales como CLIP y aumentos cercanos al 20 % en OA. Estos resultados sugieren que las huellas térmicas contienen información útil sobre interacciones recientes y ayudan a reducir la ambigüedad durante la reconstrucción.

El impacto de la modalidad térmica también cambia según el tipo de interacción. Las mejoras fueron más notorias en escenas con una transferencia importante de calor, como las interacciones de sentado. En contactos breves o con huellas térmicas débiles, la mejora fue menor. Esto indica que la calidad final depende tanto de la modalidad utilizada como de las características físicas de la interacción.

Las diferencias más grandes aparecen cuando se agregan descripciones semánticas. En esta configuración, el modelo recibe información explícita sobre el evento ocurrido antes de la captura, facilitando la reconstrucción de la escena. El puntaje CLIP alcanza valores cercanos a 93.988, mostrando una alta relación semántica entre la reconstrucción y la imagen de referencia.

Cuando la descripción utiliza el formato SPED, el desempeño vuelve a mejorar. La estructura de este esquema ayuda a reducir ambigüedades y orienta de mejor manera al modelo generativo. Este efecto se observa principalmente en las métricas de alto nivel, donde se preserva mejor la identidad de los objetos y su relación espacial.

Las métricas de bajo nivel, como PSNR, permanecen relativamente bajas (por debajo de 20). Esto se debe a que son sensibles a pequeñas variaciones en la geometría, la perspectiva o la posición de los objetos. Por el contrario, métricas como LPIPS, CLIP, OA e IoU ofrecen una visión más representativa del desempeño, ya que evalúan similitud perceptual, contenido semántico y coherencia estructural.

En conjunto, los resultados indican que la combinación de evidencia térmica y descripciones semánticas proporciona una representación más informativa del evento pasado y favorece la reconstrucción del estado previo de la escena.

La Figura 11 muestra una comparación visual de las distintas configuraciones evaluadas. Los paneles (a) y (b) corresponden a las modalidades de entrada RGB y térmica, respectivamente. El panel (c) presenta las estrategias de descripción utilizadas para guiar

el proceso de generación. La imagen de referencia se muestra en (d), mientras que las reconstrucciones obtenidas se presentan en los paneles (e) a (h).

Cuando solo se utiliza la imagen RGB, el modelo no dispone de información sobre eventos previos y tiende a conservar el estado actual de la escena. La incorporación de la modalidad térmica permite recuperar parte de la región donde ocurrió la interacción, aunque la reconstrucción sigue siendo limitada en algunos casos.

Las reconstrucciones más consistentes se obtienen al incorporar información semántica. Este efecto es aún más evidente cuando se utiliza el formato SPED, ya que la descripción estructurada proporciona una guía más clara sobre el evento ocurrido. Como resultado, las reconstrucciones presentan una mejor correspondencia con la escena de referencia y una representación más coherente del estado pasado.

5.3.4. Modelos generativos guiados por VLM El desempeño del modelo generativo es un factor determinante en la calidad perceptual y la precisión de las reconstrucciones de estados pasados. Más allá de predecir correctamente la configuración previa de la escena, el modelo debe generar una representación visual coherente, físicamente plausible y alineada con la observación RGB actual, interpretando las huellas térmicas como evidencia de interacciones anteriores.

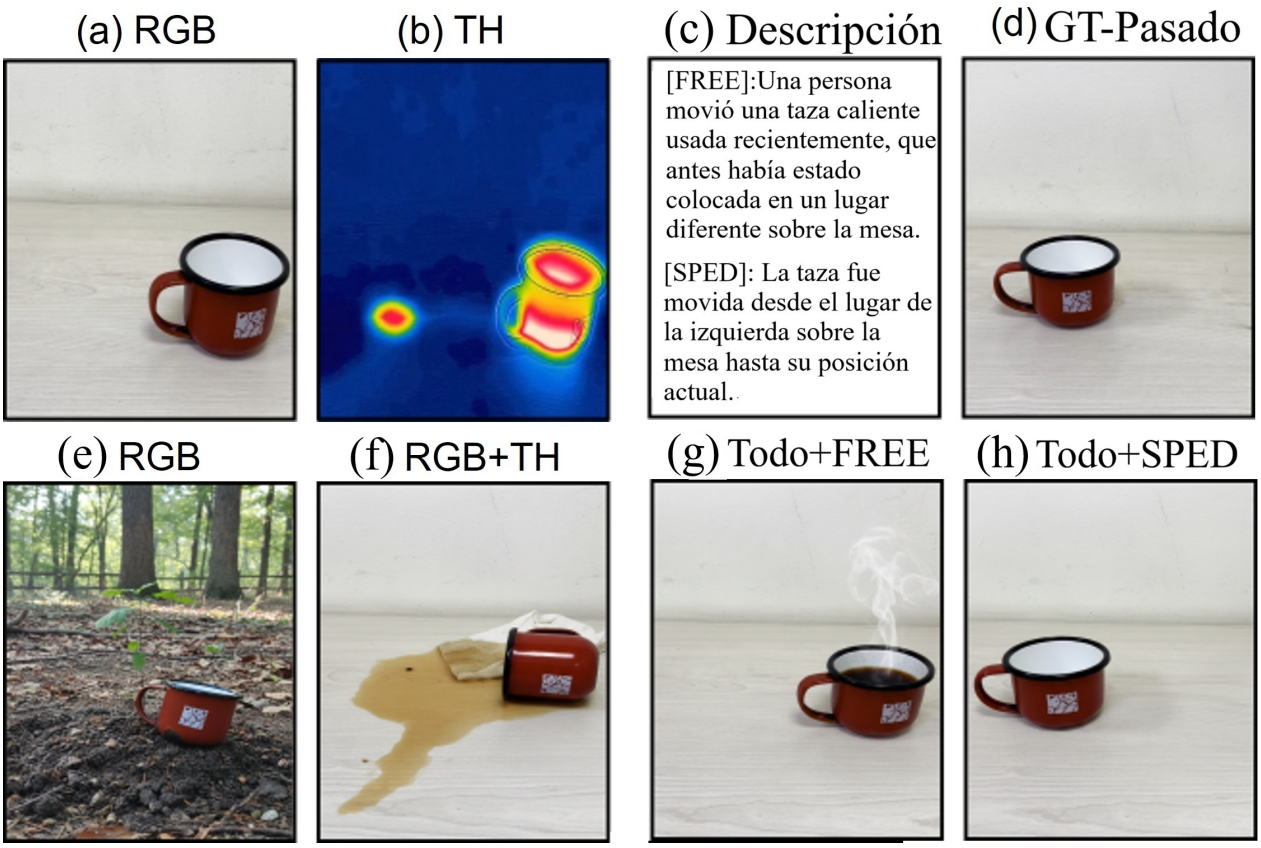
Para evaluar esta capacidad, se fija un horizonte de reconstrucción de $\Delta = 60$ segundos, el cual representa un escenario desafiante debido a la degradación progresiva de las señales térmicas con el tiempo. Este análisis permite comparar la capacidad de distintos modelos para reconstruir escenas bajo condiciones de información física limitada.

Cuadro 6. Comparación de modelos generativos para reconstrucción temporal inversa con un retardo de 60 s.

Método	**Bajo Nivel**		**Nivel de Características**		**Alto Nivel**	
	PSNR ↑	**SSIM** ↑	**LPIPS** ↓	**CLIP Sim.** ↑	**OA** ↑	**IoU** ↑
Flux Kontext 4.0	13.782 ± 5.053	0.512 ± 0.102	0.508 ± 0.265	90.189 ± 5.866	0.589 ± 0.021	0.546 ± 0.039
Gemini 2.5	**16.001 ± 1.563**	**0.615 ± 0.073**	**0.335 ± 0.086**	92.657 ± 0.658	**0.912 ± 0.069**	**0.851 ± 0.096**
DALL·E 3	13.751 ± 2.507	0.515 ± 0.094	0.417 ± 0.138	90.848 ± 0.720	0.825 ± 0.127	0.785 ± 0.135
Grok	13.389 ± 2.981	0.575 ± 0.197	0.511 ± 0.154	84.722 ± 7.896	0.633 ± 0.108	0.544 ± 0.017
Qwen-Image	14.240 ± 4.333	0.448 ± 0.162	0.462 ± 0.189	80.295 ± 13.481	0.831 ± 0.149	0.800 ± 0.147
SeedDream 4.0	15.740 ± 1.036	0.538 ± 0.047	0.338 ± 0.076	**92.882 ± 1.181**	0.879 ± 0.031	0.830 ± 0.061

Fuente: Elaboración propia a partir de los resultados experimentales.

Figura 11. **Resultados visuales del estudio de ablación.** (a–b) Modalidades de entrada: RGB y térmica (TH). (c) Descripción semántica estimada $\hat{y}_{t-\Delta}$ utilizando estrategias Libre y SPED. (d) Imagen de referencia (GT) del estado pasado. (e–h) Reconstrucciones bajo diferentes configuraciones, evidenciando el impacto de la modalidad térmica y del uso de descripciones semánticas en la calidad de la reconstrucción.



Fuente: Elaboración propia.

Se evaluaron múltiples modelos generativos guiados por modelos visión-lenguaje bajo la misma configuración experimental (RGB + TH + SPED), incluyendo DALL·E 3[50], Grok[51],

[50] OPENAI, *DALL·E 3*, ver n. 48.

[51] XAI. *Grok*. `https://x.ai`. [Fecha de consulta: 30 de septiembre de 2025]. 2025.

Qwen-Image[52], Flux Kontext 4.0[53], SeedDream 4.0[54] y Gemini 2.5[55]. Los resultados cuantitativos se presentan en el Cuadro 6.

Las métricas de bajo nivel (PSNR, SSIM) presentan valores relativamente bajos en todos los modelos. Esto se debe a su alta sensibilidad a pequeñas variaciones en la posición de objetos, cambios de perspectiva o diferencias en la geometría de la escena. En contraste, las métricas de nivel de características (LPIPS, CLIP) y de alto nivel (OA, IoU) proporcionan una evaluación más robusta de la calidad de la reconstrucción, ya que capturan similitudes perceptuales y coherencia estructural.

En este contexto, Gemini 2.5 y SeedDream 4.0 muestran un desempeño más consistente en comparación con los demás modelos. Sin embargo, Gemini 2.5 destaca particularmente en las métricas de alto nivel, lo que indica una mejor preservación de la identidad de los objetos y de las relaciones espaciales dentro de la escena.

La Figura 12 ilustra las diferencias cualitativas entre los modelos evaluados. Todos los modelos reciben las mismas entradas (RGB y térmica) junto con la misma descripción semántica estructurada. A pesar de estas condiciones controladas, se observan diferencias significativas en la calidad de las reconstrucciones.

Grok presentó varios problemas al reconstruir figuras humanas, especialmente en los rostros, donde la identidad facial cambiaba entre generaciones y producía resultados poco consistentes. En el caso de Qwen-Image, algunas reconstrucciones omitieron objetos secundarios importantes, dejando escenas incompletas.

Flux Kontext 4.0 mantuvo gran parte de la estructura general de la escena, aunque en varias imágenes aparecieron errores en la posición de los sujetos y objetos. Esto generó reconstrucciones visualmente plausibles, pero con inconsistencias respecto a la interacción original.

---

[52] PIXVERSE. *PixVerse AI Video Generator*. `https://app.pixverse.ai`. [Fecha de consulta: 30 de septiembre de 2025]. 2025.

[53] Stephen BATIFOL *et al.* «FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space». En: *arXiv e-prints* (2025), arXiv-2506.

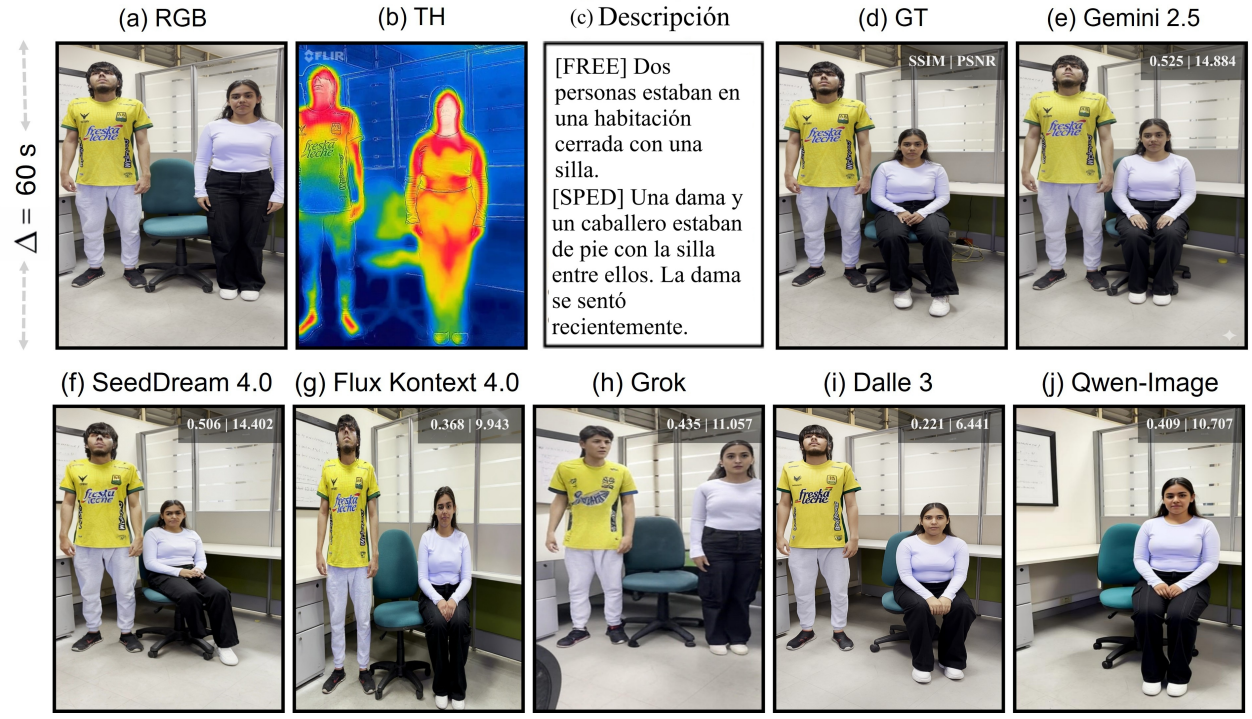
[54] Team SEEDREAM *et al.* «Seedream 4.0: Toward Next-generation Multimodal Image Generation». En: *arXiv preprint arXiv:2509.20427* (2025).

[55] GOOGLE, ver n. 49.

SeedDream 4.0 logró conservar el contexto principal de la escena, aunque en algunos casos modificó la perspectiva de la imagen, produciendo efectos similares a un *zoom-out*. Estas variaciones afectaron la similitud directa con la imagen de referencia. Por otro lado, DALL·E 3 generó imágenes coherentes desde el punto de vista visual, pero con un exceso de suavizado y cambios en la distribución de algunos elementos de la escena. Entre los modelos evaluados, Gemini 2.5 produjo las reconstrucciones más cercanas a la referencia original (GT). En la mayoría de los casos, logró mantener tanto la distribución espacial de la escena como la interacción principal observada.

5.3.5. Rango temporal de reconstrucción El experimento final evalúa hasta qué punto en el pasado es posible reconstruir de manera confiable el estado de una escena a par-

Figura 12. **Comparación de modelos generativos guiados por VLM para reconstrucción temporal.** Cada modelo recibe las mismas entradas (RGB y térmica) y la misma descripción semántica. Gemini 2.5 produce las reconstrucciones más cercanas a la verdad de referencia (GT), preservando la estructura espacial y la coherencia semántica de la escena.



Fuente: Elaboración propia.

tir de huellas residuales. Para este análisis, se generaron reconstrucciones considerando retardos temporales crecientes (5 s, 15 s, 30 s, 60 s y 120 s), comparando cada resultado con la correspondiente verdad de referencia (GT). Todas las reconstrucciones se obtuvieron utilizando la configuración completa del sistema (RGB + TH + SPED) y el mismo modelo generativo.

Cuadro 7. **Rango de reconstrucción temporal.** Desempeño a medida que aumenta el retardo de reconstrucción, reflejando la degradación progresiva de las trazas residuales.

Delay (s)	**Bajo Nivel**		**Nivel de características**		**Alto nivel**	
	PSNR ↑	**SSIM** ↑	**LPIPS** ↓	**CLIP** ↑	**OA** ↑	**IoU** ↑
5	18.551 $\pm$ 2.342	0.776 $\pm$ 0.114	0.325 $\pm$ 0.126	91.932 $\pm$ 3.919	0.919 $\pm$ 0.120	0.887 $\pm$ 0.116
15	18.549 $\pm$ 3.000	0.775 $\pm$ 0.106	0.321 $\pm$ 0.099	91.984 $\pm$ 1.674	0.918 $\pm$ 0.129	0.873 $\pm$ 0.133
30	18.268 $\pm$ 2.733	0.718 $\pm$ 0.146	0.297 $\pm$ 0.195	92.755 $\pm$ 3.177	0.854 $\pm$ 0.249	0.739 $\pm$ 0.177
60	16.001 $\pm$ 1.563	0.615 $\pm$ 0.073	0.335 $\pm$ 0.086	92.657 $\pm$ 0.658	0.912 $\pm$ 0.069	0.851 $\pm$ 0.096
120	14.668 $\pm$ 2.355	0.631 $\pm$ 0.105	0.325 $\pm$ 0.105	86.716 $\pm$ 4.307	0.774 $\pm$ 0.123	0.711 $\pm$ 0.164

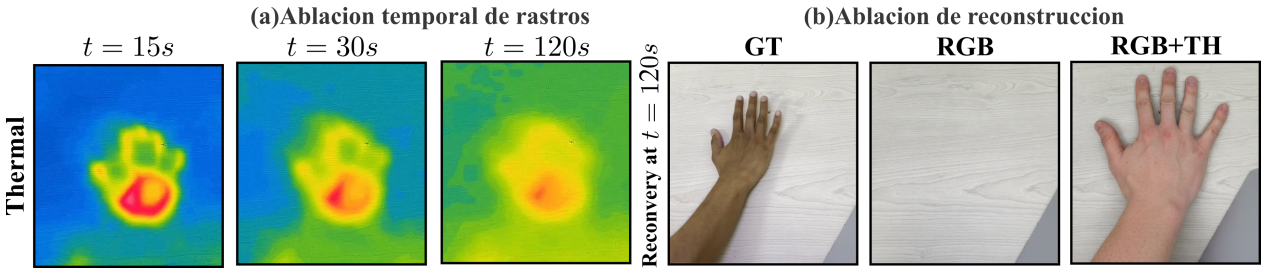
Fuente: Elaboración propia a partir de los resultados experimentales.

El Cuadro 7 presenta los resultados cuantitativos para los diferentes retardos. Se observa que el desempeño se mantiene relativamente estable hasta aproximadamente 30 segundos. A partir de este punto, comienza una degradación progresiva, que se hace más evidente a los 60 segundos y se intensifica significativamente a partir de los 120 segundos, donde las reconstrucciones pierden precisión tanto a nivel estructural como semántico.

Las métricas de alto nivel (OA e IoU) muestran una mayor robustez en comparación con las métricas de bajo nivel, manteniéndose relativamente estables hasta aproximadamente 60 segundos. Esto indica que, aunque la calidad pixel a pixel disminuye, el sistema aún logra preservar la identidad de los objetos y sus relaciones espaciales dentro de este rango temporal. Sin embargo, más allá de este punto, la pérdida de información física incrementa la incertidumbre en la reconstrucción, afectando también la consistencia semántica.

La Figura 13(**a**) muestra la evolución de las huellas térmicas a lo largo del tiempo. Inmediatamente después de la interacción, todavía es posible distinguir la estructura de contacto, como la forma de una mano. Sin embargo, la señal disminuye progresivamente debido a la difusión del calor y, después de aproximadamente 120 segundos, gran parte de esta información deja de ser visible. Esta pérdida de información también se refleja en

Figura 13. **Impacto de la degradación temporal de huellas térmicas en la reconstrucción visual.** (**a**) Disipación progresiva de las huellas térmicas con el paso del tiempo. (**b**) Comparación de reconstrucciones para retardos largos. RGB-only conserva el estado actual de la escena y RGB+TH produce resultados más cercanos a la referencia (GT).



Fuente: Elaboración propia.

las métricas, las cuales disminuyen a medida que aumenta el retardo temporal.

La Figura 13(**b**) muestra ejemplos de reconstrucción para un retardo de 120 segundos. Cuando el sistema utiliza solo la imagen RGB, normalmente conserva el estado actual de la escena, ya que no tiene evidencia sobre interacciones anteriores. Al agregar información térmica (RGB+TH), el modelo logra recuperar parte de la interacción; sin embargo, a este nivel de retardo la señal térmica ya es bastante débil y las reconstrucciones pueden volverse menos precisas.

La configuración completa que incorpora SPED genera reconstrucciones más cercanas al evento original. Incluso cuando la huella térmica es limitada, la descripción semántica aporta contexto adicional y ayuda al modelo durante la reconstrucción.

En conjunto, los resultados muestran que el límite temporal del método depende principalmente de cuánto tiempo permanezcan visibles las huellas térmicas. También indican que la información semántica ayuda a compensar parcialmente la pérdida de información física cuando aumenta el tiempo entre el evento y la captura.

En general, los experimentos realizados permitieron analizar el comportamiento del sistema bajo distintas configuraciones, modelos generativos y retardos temporales. El uso de métricas de bajo nivel, perceptuales y semánticas permitió evaluar tanto la calidad visual de las reconstrucciones como su relación con la escena original.

Los resultados obtenidos muestran ventajas y limitaciones del método propuesto, espe-

cialmente cuando disminuye la información térmica disponible. De esta manera, esta sección da cumplimiento al Objetivo Específico 4, al evaluar el desempeño del sistema mediante comparaciones cuantitativas y cualitativas frente a datos reales.

5.4. REPOSITORIOS DE REFERENCIA

Para facilitar la reproducción de los experimentos, se organizaron repositorios digitales con el código y los datos utilizados durante la investigación. Estos repositorios permitieron mantener un registro de los experimentos y documentar los cambios realizados durante el desarrollo del sistema.

5.4.1. Repositorio en GitHub El repositorio alojado en GitHub[56] constituye el núcleo del desarrollo metodológico del presente trabajo, ya que contiene la implementación completa del framework propuesto para la reconstrucción temporal a partir de señales térmicas. Este repositorio fue diseñado bajo principios de modularidad, trazabilidad y documentación clara, con el fin de facilitar su uso, comprensión y extensión por parte de otros investigadores.
En particular, el repositorio incluye:

- **Framework de evaluación**: El directorio `src/` contiene los módulos utilizados para calcular métricas como PSNR, SSIM, IoU, LPIPS y CLIP.
- **Procesamiento experimental**: Se incluyen scripts para cargar los datos, ejecutar las pruebas y generar los resultados. Entre ellos se encuentra `generate_experimental_results` utilizado durante las evaluaciones principales.
- **Modelos integrados**: El sistema utiliza herramientas como YOLOv11 para detección y segmentación, y CLIP (ViT-B/32) para comparar similitud semántica entre imágenes.
- **Organización de datos**: El repositorio contiene las imágenes de referencia (*ground*

[56] Luis TOSCANO. *Trabajo de grado*. `https://github.com/JustBeingLuis/Trabajo-de-grado`. Repositorio de código fuente del proyecto. [Fecha de consulta: 9 de abril de 2026]. 2026.

truth), las reconstrucciones generadas y los resultados obtenidos en los experimentos.

- **Generación de reportes**: El sistema genera archivos CSV y resúmenes en texto que permiten consolidar y analizar los resultados obtenidos durante los experimentos.

Además, el repositorio incluye documentación para la instalación y configuración del entorno de trabajo, junto con la especificación de dependencias y requisitos computacionales. El uso de GitHub facilita el control de versiones y el seguimiento de cambios realizados durante el desarrollo del proyecto.

5.4.2. Repositorio en Kaggle El conjunto de datos empleado en este trabajo se encuentra disponible públicamente en Kaggle[57]. Este recurso reúne la base experimental del proyecto en un entorno abierto y fácil de consultar,y obviamente facilita el acceso a las muestras multimodales utilizadas en la investigación y fortalece la reproducibilidad del trabajo. Esto resulta útil porque concentra en un mismo lugar las observaciones RGB y térmicas de las escenas, manteniendo una organización adecuada para su descarga, revisión y reutilización en tareas de validación o comparación experimental.
El repositorio en Kaggle ayuda que otros investigadores exploren el conjunto de datos sin depender del entorno local en el que se desarrolló la tesis. Esto hace más sencillo repetir experimentos, probar variantes del flujo propuesto o incluso usar la base en estudios posteriores sobre inferencia temporal, visión térmica o reconstrucción guiada por lenguaje.El repositorio complementa el cumplimiento del Objetivo Específico 4, al dejar disponible la evidencia experimental sobre la cual se apoyan los análisis y resultados presentados en este documento.

---

[57] Luis TOSCANO. *Repositorio de datos en Kaggle*. `https://www.kaggle.com/datasets/ludwig645/traceeeeee`. Repositorio de datasets y resultados experimentales. [Fecha de consulta: 9 de abril de 2026]. 2026.

6. CONCLUSIONES

En este trabajo se validó el paradigma de *reconstrucción temporal inversa*, orientado a inferir el estado anterior de una escena mediante la codificación pasiva de huellas térmicas. Al finalizar la investigación, se da respuesta a la pregunta de investigación mediante las siguientes conclusiones:

Se demostró que el diseño de un algoritmo de reconstrucción efectivo requiere una arquitectura híbrida que combine el sensado físico con el razonamiento semántico de alto nivel. La información térmica, aunque crítica para reducir la ambigüedad del problema inverso, actúa solo como una guía que debe ser interpretada. En este sentido, el formato *SPED* se consolidó como el mecanismo de control necesario para traducir señales físicas en instrucciones visuales coherentes.

Los experimentos realizados mostraron que el tamaño del modelo influye directamente en la calidad de las reconstrucciones. Los modelos ligeros ejecutados localmente presentaron dificultades para interpretar correctamente las huellas térmicas producidas por interacciones humanas. En cambio, modelos de mayor escala como **GPT-5** generaron descripciones más claras y reconstrucciones más cercanas a la escena original.

Las métricas obtenidas también muestran que la calidad de la reconstrucción no puede evaluarse únicamente mediante diferencias de píxel. Aunque métricas como PSNR o SSIM permiten medir similitud visual, no siempre reflejan correctamente el contenido de la escena. Durante los experimentos, los modelos de visión y lenguaje de gran escala mantuvieron una identificación adecuada de objetos incluso cuando la señal térmica era débil, particularmente en retardos cercanos a los 120 segundos.

En general, los resultados indican que la combinación de evidencia térmica residual y modelos generativos multimodales permite recuperar información relevante sobre eventos recientes. Además, el conjunto de datos *TRACE-HEI* y la metodología desarrollada pueden servir como base para futuras investigaciones sobre reconstrucción retrospectiva mediante huellas térmicas.

7. TRABAJO FUTURO

Los resultados obtenidos también muestran varias posibilidades para continuar esta línea de trabajo.

Una de ellas consiste en evaluar el sistema en escenarios reales y no únicamente en ambientes controlados. Para ello, sería necesario trabajar con sistemas de captura sincronizada capaces de operar bajo cambios de temperatura, iluminación y ruido térmico ambiental.

Otra línea de interés está relacionada con la ejecución local de los modelos. Debido a las limitaciones observadas en las arquitecturas ligeras, resulta relevante explorar técnicas que permitan mejorar su desempeño sin depender completamente de infraestructura externa.

Finalmente, también sería útil incorporar variables físicas como la emisividad, la conductividad y la difusividad térmica dentro del proceso de reconstrucción. Esto permitiría modelar de mejor manera la evolución de las huellas térmicas y estudiar su efecto sobre el tiempo máximo de reconstrucción.

De igual forma, la codificación pasiva podría explorarse en otros espectros, como el ultravioleta (UV), con el fin de aprovechar evidencias residuales que presentan dinámicas temporales diferentes a las señales térmicas.

Finalmente, sería conveniente evaluar el método en escenarios más complejos, con múltiples personas y objetos interactuando simultáneamente, para analizar sus capacidades y limitaciones en condiciones más cercanas a aplicaciones reales.

BIBLIOGRAFÍA

ABDELRAHMAN, Yomna *et al.* «Stay Cool! Understanding Thermal Attacks on Mobile-based User Authentication». En: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI)*. 2017, pp. 3751-3763 (vid. pág. 18).

ANTHROPIC. *Claude 3.5: Next-generation AI assistant*. https://www.anthropic.com/news/claude-3-5-sonnet. [Fecha de consulta: 28 de marzo de 2026]. 2024 (vid. pág. 41).

BATIFOL, Stephen *et al.* «FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space». En: *arXiv e-prints* (2025), arXiv-2506 (vid. pág. 52).

BERTERO, Mario y Patrizia BOCCACCI. *Introduction to Inverse Problems in Imaging*. Institute of Physics Publishing, 1998 (vid. págs. 18, 28).

BROOKS, Tom, Anna HOLYNSKI y Alex A. EFROS. «InstructPix2Pix: Learning to follow image editing instructions». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 18392-18402. DOI: 10.1109/CVPR52729.2023.01764 (vid. págs. 25, 28).

BROWN, Tom B. *et al.* «Language Models are Few-Shot Learners». En: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020 (vid. pág. 24).

CARDONE, Daniela y Arcangelo MERLA. «New frontiers for applications of thermal infrared imaging devices». En: *Sensors* 17.5 (2017), pág. 1042 (vid. pág. 16).

CÓRDOBA-TLAXCALTECO, M. L. y Edgard BENÍTEZ-GUERRERO. «Human event recognition in smart classrooms using computer vision: a systematic literature review». En: *Programming and Computer Software* 49.8 (2023), pp. 625-642 (vid. pág. 14).

EVERINGHAM, Mark *et al.* «The Pascal Visual Object Classes (VOC) challenge». En: *International Journal of Computer Vision* 88.2 (2010), pp. 303-338 (vid. pág. 40).

FLORES, Carlos y Santiago GONZÁLEZ. «Evaluación y Comparación de Métricas Objetivas PSNR, SSIM y LPIPS para el Análisis de Calidad de Video». En: *Revista Tecnológica - ESPOL* 37.E1 (2025), pág. 1317. DOI: 10.37815/rte.v37nE1.1317 (vid. pág. 25).

GAO, Hang *et al.* «Disentangling propagation and generation for video prediction». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 9006-9015 (vid. pág. 15).

GOOGLE. *Introducing Gemini 2.5 Flash Image (aka Nano Banana)*. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/. [Fecha de consulta: 30 de septiembre de 2025]. 2025 (vid. págs. 48, 52).

GOOGLE DEEPMIND. *Gemini 3: Next-Generation Multimodal Models*. https://deepmind.google/technologies/gemini/. [Fecha de consulta: 10 de abril de 2026]. 2026 (vid. pág. 41).

HERTZ, Amir *et al.* «Prompt-to-Prompt Image Editing with Cross-Attention Control». En: *arXiv preprint arXiv:2208.01626* (2022) (vid. pág. 24).

HU, Edward J. *et al.* «LoRA: Low-Rank Adaptation of Large Language Models». En: *International Conference on Learning Representations*. 2022 (vid. págs. 30, 32).

HUGGING FACE. *SmolVLM-Instruct*. https://huggingface.co/HuggingFaceTB/SmolVLM-Instruct. [Model card en Hugging Face. Fecha de consulta: 5 de abril de 2026]. 2025 (vid. pág. 32).

JOCHER, Glenn *et al. Ultralytics YOLOv11*. https://github.com/ultralytics/ultralytics. 2024 (vid. pág. 40).

KETSEKIOULAFIS, Ioannis *et al.* «Artificial Intelligence in Forensic Sciences: a systematic review of past and current applications and future perspectives». En: *Cureus* 16.9 (2024) (vid. pág. 14).

KIRILLOV, Alexander, Eric MINTUN, Nikhila RAVI *et al.* «Segment Anything». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 4015-4026 (vid. pág. 40).

KOESDWIADY, Arief *et al.* «Recent trends in driver safety monitoring systems: State of the art and challenges». En: *IEEE Transactions on Vehicular Technology* 66.6 (2016), pp. 4550-4563 (vid. pág. 14).

LARSON, Eric *et al.* «HeatWave: Thermal imaging for surface user interaction». En: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2011. DOI: 10.1145/1978942.1979317 (vid. pág. 21).

LI, Junnan *et al.* «BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation». En: *Proceedings of the International Conference on Machine Learning (ICML)*. 2022 (vid. pág. 24).

LIU, Haotian *et al.* «Improved baselines with visual instruction tuning». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 26296-26306 (vid. pág. 16).

LIU, Haotian *et al. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge*. https://llava-vl.github.io/blog/2024-01-30-llava-next/. [Fecha de consulta: 10 de abril de 2026]. 2024 (vid. pág. 41).

LIU, Zhengqiang *et al.* «Video frame synthesis using deep voxel flow». En: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017 (vid. pág. 22).

LIU, Ziwei *et al.* «Video frame synthesis using deep voxel flow». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2017, pp. 4463-4471 (vid. pág. 15).

LOTTER, William, Gabriel KREIMAN y David COX. «Deep predictive coding networks for video prediction and unsupervised learning». En: *arXiv preprint arXiv:1605.08104* (2016) (vid. págs. 15, 22).

MAZUR-MILECKA, M. *et al.* «Detection and Model of Thermal Traces Left after Aggressive Behavior of Laboratory Rodents». En: *Applied Sciences* 11.14 (2021), pág. 6644 (vid. pág. 18).

OPENAI. *DALL·E 3*. `https://openai.com/dall-e-3`. [Fecha de consulta: 30 de septiembre de 2025]. 2023 (vid. págs. 48, 51).

— *GPT-5 Technical Report*. `https://openai.com/research/gpt-5`. [Fecha de consulta: 10 de abril de 2026]. 2025 (vid. pág. 41).

PIXVERSE. *PixVerse AI Video Generator*. `https://app.pixverse.ai`. [Fecha de consulta: 30 de septiembre de 2025]. 2025 (vid. pág. 52).

QWEN TEAM *et al.* «Qwen3.5: Advancing Large Vision-Language Models». En: *arXiv preprint* (2025) (vid. pág. 41).

RADFORD, Alec *et al.* «Learning transferable visual models from natural language supervision». En: *Proceedings of the International Conference on Machine Learning (ICML)*. 2021, pp. 8748-8763 (vid. págs. 16, 39).

ROBINSON, Nicole *et al.* «Robotic vision for human-robot interaction and collaboration: A survey and systematic review». En: *ACM Transactions on Human-Robot Interaction* 12.1 (2023) (vid. pág. 14).

ROHERA, Pritika *et al.* «Better To Ask in English? Evaluating Factual Accuracy of Multilingual LLMs in English and Low-Resource Languages». En: *arXiv preprint arXiv:2504.20022* (2025) (vid. pág. 27).

ROMBACH, Robin *et al.* «High-resolution image synthesis with latent diffusion models». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10684-10695 (vid. págs. 16, 23).

SAHARIA, Chitwan *et al.* «Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding». En: *arXiv preprint arXiv:2205.11487* (2022) (vid. pág. 24).

SEEDREAM, Team *et al.* «Seedream 4.0: Toward Next-generation Multimodal Image Generation». En: *arXiv preprint arXiv:2509.20427* (2025) (vid. pág. 52).

TAHA, Abdel Aziz y Allan HANBURY. «Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool». En: *BMC Medical Imaging* 15.1 (2015), pág. 29 (vid. pág. 40).

TANG, Zitian *et al.* «What happened 3 seconds ago? Inferring the past with thermal imaging». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 17111-17120 (vid. págs. 16, 19, 28).

TOSCANO, Luis. *Repositorio de datos en Kaggle*. `https://www.kaggle.com/datasets/ludwig645/traceeeeee`. Repositorio de datasets y resultados experimentales. [Fecha de consulta: 9 de abril de 2026]. 2026 (vid. pág. 57).

— *Trabajo de grado*. `https://github.com/JustBeingLuis/Trabajo-de-grado`. Repositorio de código fuente del proyecto. [Fecha de consulta: 9 de abril de 2026]. 2026 (vid. pág. 56).

VOGEL, Curtis R. *Computational Methods for Inverse Problems*. Society for Industrial y Applied Mathematics, 2002. DOI: 10.1137/1.9780898717570 (vid. pág. 21).

VOLLMER, Michael y Klaus-Peter MÖLLMANN. *Infrared thermal imaging: fundamentals, research and applications*. John Wiley & Sons, 2018 (vid. pág. 16).

WANG, Peng *et al.* «Qwen2-VL: Enhancing vision-language model's perception». En: *arXiv preprint arXiv:2409.12191* (2024) (vid. pág. 16).

WANG, Zhou *et al.* «Image quality assessment: From error visibility to structural similarity». En: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600-612 (vid. pág. 39).

XAI. *Grok*. https://x.ai. [Fecha de consulta: 30 de septiembre de 2025]. 2025 (vid. pág. 51).

YANG, An *et al.* «Qwen2 Technical Report». En: *arXiv preprint arXiv:2407.10671* (2024) (vid. pág. 30).

YU, Wei *et al.* «Crevnet: Conditionally reversible video prediction». En: *arXiv preprint arXiv:1910.11577* (2019) (vid. págs. 15, 22).

ZHANG, Lingzhi, Aditya RAO y Maneesh AGRAWALA. «Adding Conditional Control to Text-to-Image Diffusion Models». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. DOI: 10.1109/ICCV51070.2023.00355 (vid. págs. 23, 28).

ZHANG, Richard *et al.* «The unreasonable effectiveness of deep features as a perceptual metric». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 586-595 (vid. pág. 39).

ANEXO A. BASE DE DATOS TRACE

TRACE-HEI es el primer conjunto de datos multimodal diseñado para la inferencia de escenas invertidas en el tiempo, capturando la decadencia temporal de señales térmicas y visibles (RGB) después de interacciones humano–entorno. A continuación se describen los esquemas de adquisición y se proporcionan estadísticas del conjunto de datos.

ESQUEMAS DE ADQUISICIÓN

Se emplearon dos cámaras:

- **Cámara térmica.** Los datos térmicos fueron registrados utilizando la FLIR ONE Pro [58], un sensor de infrarrojo de onda larga con una resolución espacial de 160×120 píxeles. Opera en la banda espectral de 8–14 μm, capturando la transferencia de calor residual dejada después del contacto humano. Esta modalidad proporciona información crucial sobre la dinámica de decaimiento de temperatura, particularmente dentro de los primeros 1–120 segundos después de la interacción.
- **Cámara RGB.** Las imágenes en el espectro visible fueron capturadas utilizando la cámara principal del iPhone 15 [59], produciendo imágenes a color con resolución de 4032×3024 dentro del rango espectral de 400–700 nm. Esta modalidad proporciona información geométrica y de apariencia de alta calidad, sirviendo como el marco de referencia que el modelo generativo debe preservar al reconstruir configuraciones pasadas plausibles.

Todos los dispositivos fueron montados sobre un soporte rígido y calibrados manualmente para asegurar la alineación espacial. Para cada interacción, el evento en sí tiene una duración entre 5 s y 30 s, y las trazas en decaimiento se registran hasta por 180 s (3 minutos), capturando diferentes etapas de disipación térmica.

---

[58] `https://www.flir.com/products/flir-one-pro/`

[59] `https://www.apple.com/iphone-15/`

FLUJO DE PROCESAMIENTO DE DATOS

Todas las grabaciones en TRACE-HEI pasan por un pipeline de post-procesamiento estandarizado para garantizar consistencia temporal, alineación espacial y supervisión multimodal de alta calidad entre las secuencias RGB y térmica.

1. Sincronización temporal. Aunque las grabaciones se inician simultáneamente, pequeñas desviaciones entre sensores son corregidas en el post-procesamiento. Sincronizamos los fotogramas emparejando señales visuales (p. ej., entrada de la mano, contacto con objetos) y verificando manualmente la alineación entre modalidades. Todas las secuencias se remuestrean a una tasa de cuadros unificada (FPS) para garantizar correspondencia fotograma a fotograma.

2. Alineación espacial. Los sensores poseen diferentes campos de visión y parámetros intrínsecos. Para obtener tripletas alineadas a nivel de píxel, (i) corregimos la distorsión de cada modalidad utilizando los parámetros de calibración del fabricante, (ii) estimamos homografías mediante alineación con tablero de ajedrez cuando es posible, y (iii) refinamos manualmente la alineación en superficies con baja estructura térmica. Los fotogramas finales alineados se recortan a la región común de todas las cámaras.

3. Normalización de intensidad térmica. Los valores térmicos del sensor FLIR presentan variabilidad debido a deriva de temperatura y condiciones de captura. Se aplica una normalización min–max por secuencia.

4. Selección de fotogramas y etapas temporales. Para cada escena, se extraen tripletas sincronizadas en instantes predefinidos: $\Delta \in \{5, 15, 30, 60, 120, 180\}$ segundos después de la interacción. Estos se utilizan como información de entrada para el proceso de inferencia temporal.

5. Consolidación de anotaciones. Las etiquetas relacionadas con la interacción, incluyendo acciones, objetos, materiales y entorno, fueron anotadas manualmente. Los metadatos de cada escena se almacenan en un archivo `.xlsx` que acompaña el conjunto de datos.

ANEXO B. ESPECIFICACIONES TÉCNICAS DEL ENTRENAMIENTO DE MODELOS

Este anexo presenta los parámetros de configuración y los hiperparámetros utilizados durante el ajuste fino de los modelos explorados en el sistema propuesto. El propósito es documentar las condiciones bajo las cuales se realizaron los experimentos en entornos con recursos de hardware limitados, así como mostrar el comportamiento del entrenamiento a partir de las curvas de pérdida y convergencia.

CONFIGURACIÓN DEL MODELO DE LENGUAJE (TEXTUAL)

El modelo textual se utiliza para transformar descripciones libres en representaciones con formato SPED. Para esta etapa se buscó una arquitectura que ofreciera un equilibrio entre capacidad de razonamiento y costo computacional.

Arquitectura base y protocolo de entrada Se seleccionó el modelo `Qwen/Qwen2.5-3B-Instruct` debido a su buen desempeño en tareas de seguimiento de instrucciones. El esquema de entrada utiliza un formato conversacional para separar las instrucciones del sistema de las entradas del usuario. Las especificaciones correspondientes se resumen en el Cuadro 8.

Cuadro 8. Especificaciones generales del modelo textual.

Componente	**Especificación Técnica**
Modelo base	`Qwen/Qwen2.5-3B-Instruct`
Parámetros totales	3 mil millones (3B)
Formato de entrada	Estructura tipo chat (Template Jinja)
Objetivo de salida	Mapeo determinista hacia la estructura SPED

Fuente: Elaboración propia.

Ajuste fino y estrategia de entrenamiento Para el ajuste del modelo se empleó la técnica LoRA, la cual permite modificar únicamente una pequeña parte de los parámetros del modelo y reducir el consumo de memoria durante el entrenamiento. Los parámetros utilizados para el adaptador y la ejecución se presentan en los cuadros 9 y 10.

Cuadro 9. Configuración del adaptador LoRA para el modelo textual.

Parámetro	**Valor**	**Descripción técnica**
`r`	16	Rango de las matrices de actualización
`lora_alpha`	32	Factor de escalamiento de la adaptación
`lora_dropout`	0.05	Coeficiente de regularización ligera
`target_modules`	*	Capas de atención y perceptrones

Fuente: Elaboración propia.

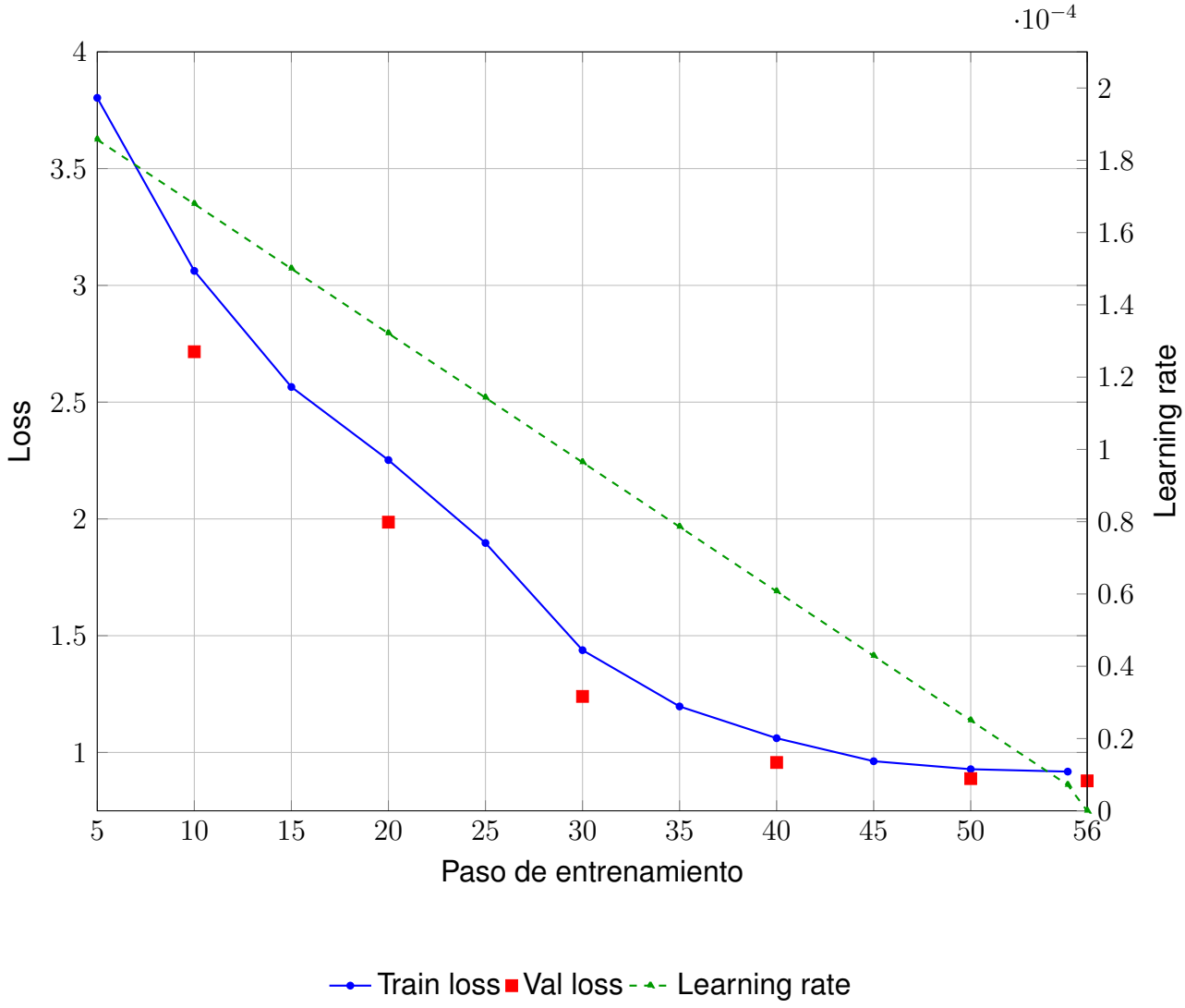
Cuadro 10. Parámetros de ejecución para el modelo textual.

Hiperparámetro	**Valor**	**Función en el entrenamiento**
`Batch size (ef.)`	8	Acumulación para gestión de memoria
`Learning rate`	`2e-4`	Tasa de aprendizaje inicial
`Epochs`	3	Ciclos de entrenamiento completos
`Optimización`	fp16	Precisión de punto flotante de 16 bits

Fuente: Elaboración propia.

La Figura 14 ilustra el proceso de convergencia del modelo textual. Se observa un descenso pronunciado de la pérdida durante los primeros 25 pasos, estabilizándose hacia el final del tercer ciclo.

Figura 14. **Evolución del entrenamiento del modelo textual SPED.** Se presentan las métricas de entrenamiento del modelo creado a partir de *Qwen2.5-3B-Instruct* para la tarea de conversión de descripciones crudas a sentencias con formato SPED. La curva evidencia una disminución sostenida de la pérdida de entrenamiento y una mejora consistente en validación a lo largo de 56 pasos globales, acompañadas por una tasa de aprendizaje decreciente hasta valores cercanos a cero.



Fuente: Elaboración propia con base en el log de entrenamiento textual.

CONFIGURACIÓN DEL MODELO VISIÓN-LENGUAJE (MULTIMODAL)

El componente multimodal integra señales RGB y térmicas para la inferencia de eventos. Esta sección detalla la configuración del VLM especializado en la detección de trazas de

interacción.

Arquitectura base y fusión sensorial Se empleó el modelo `SmolVLM-500M-Instruct`, cuyas especificaciones se presentan en el Cuadro 11.

Cuadro 11. Especificaciones generales del modelo multimodal.

Componente	**Especificación Técnica**
Modelo base	`SmolVLM-500M-Instruct`
Entradas visuales	Par sincronizado RGB + Térmico
Resolución máx.	256 $\times$ 256 píxeles (Filtro LANCZOS)
Salida esperada	Oración SPED condicionada

Fuente: Elaboración propia.

Estrategia de entrenamiento multimodal Los parámetros de optimización se resumen en los Cuadros 12 y 13.

Cuadro 12. Configuración del adaptador LoRA para el modelo multimodal.

Parámetro	**Valor**	**Descripción técnica**
`lora_r`	4	Rango reducido para estabilidad
`lora_alpha`	8	Factor de escalamiento
`lora_dropout`	0.05	Coeficiente de regularización

Fuente: Elaboración propia.

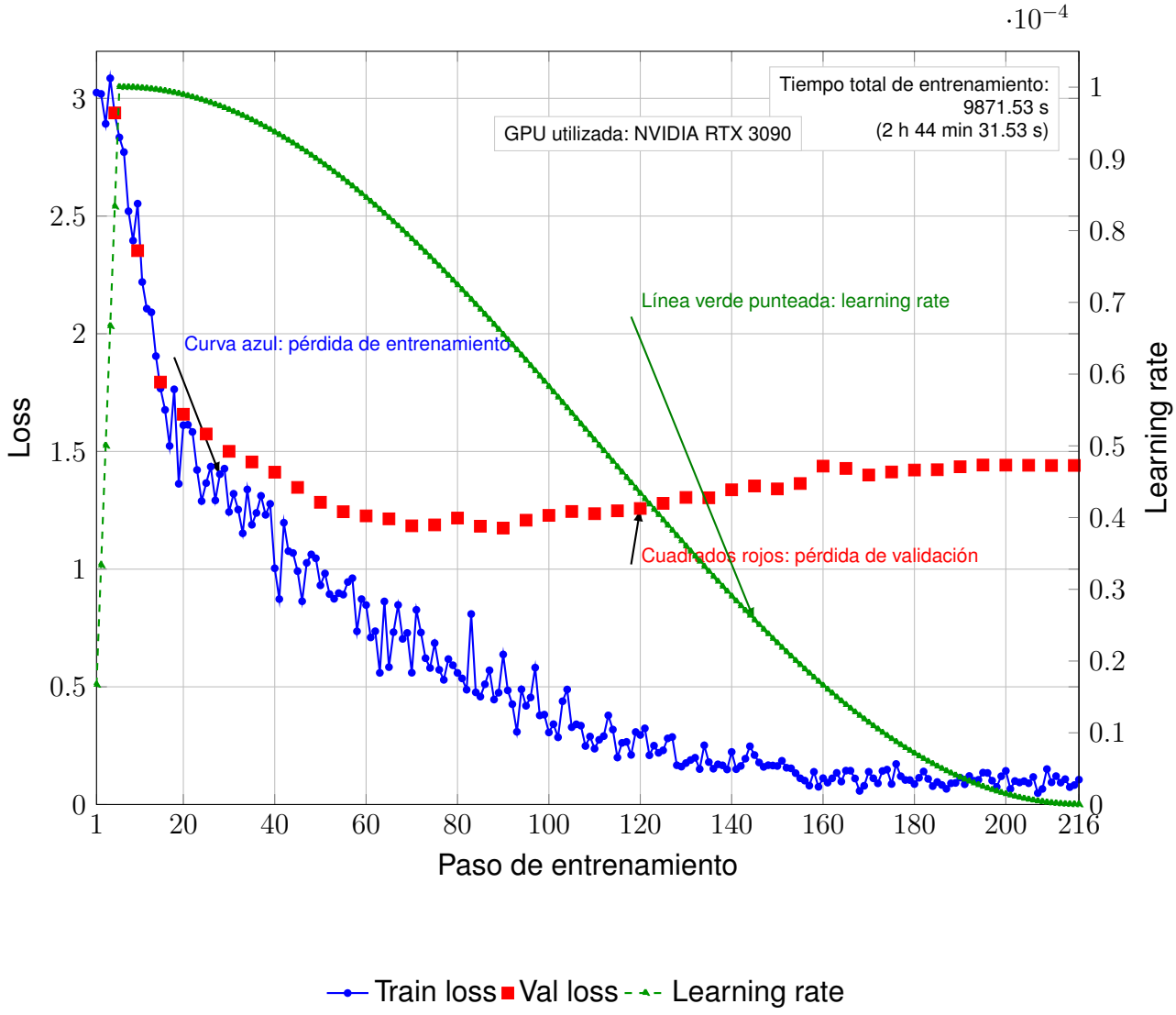
Cuadro 13. Parámetros de ejecución para el modelo multimodal.

Hiperparámetro	**Valor**	**Función en el entrenamiento**
`Batch size (ef.)`	16	Acumulación de gradientes
`Learning rate`	`2e-4`	Tasa de aprendizaje programada
`Warmup ratio`	0.03	Incremento gradual inicial
`Sampling`	`False`	Inferencia determinista

Fuente: Elaboración propia.

La Figura 15 detalla la trayectoria de optimización a lo largo de 216 pasos globales.

Figura 15. **Evolución del entrenamiento del modelo multimodal.** Se presenta la evolución de la pérdida de entrenamiento (*train loss*), la pérdida de validación (*val loss*) y la tasa de aprendizaje (*learning rate*) a lo largo de los 216 pasos de optimización. Además, se indica el tiempo total de entrenamiento registrado en el log, correspondiente a **9871.53 s** (**2 h 44 min 31.53 s**) sobre una **NVIDIA RTX 3090**. La figura permite distinguir visualmente cada curva y sus puntos de referencia más relevantes.



Fuente: Elaboración propia a partir del log de entrenamiento del modelo multimodal.