SEE THE PAST: TIME-REVERSED SCENE RECONSTRUCTION FROM THERMAL TRACES USING VISUAL LANGUAGE MODELS

Kebin Contreras[1], *Luis Toscano-Palomino*[2], *Mauro Dalla Mura*[3,4] *and Jorge Bacca*[2]

[1] Physics School, [2]Department of Computer Science, Universidad Industrial de Santander, Colombia.
[3]GIPSA-Lab, Université Grenoble Alpes, CNRS, Grenoble INP, Grenoble, France.
[4]Institut Universitaire de France (IUF), France.
jbacquin@uis.edu.co

ABSTRACT

Recovering the past from present observations is an intriguing challenge with potential applications in forensics and scene analysis. Thermal imaging, operating in the infrared range, provides access to otherwise invisible information. Since humans are typically warmer ($\approx$ 37 °C / 98.6 °F) than their surroundings, interactions such as sitting, touching, or leaning leave residual heat traces. These fading imprints serve as passive temporal codes, allowing for the inference of recent events that exceed the capabilities of RGB cameras. This work proposes a time-reversed reconstruction framework that uses paired RGB and thermal images to recover scene states from a few seconds earlier. The proposed approach couples Visual–Language Models (VLMs) with a constrained diffusion process, where one VLM generates scene descriptions and another guides image reconstruction, ensuring semantic and structural consistency. The method is evaluated in three controlled scenarios, demonstrating the feasibility of reconstructing plausible past frames up to 120 seconds earlier, providing a first step toward time-reversed imaging from thermal traces.

Index Terms— Time-reversed imaging, Thermal Imaging, Computational Imaging, Visual Language Models

1. INTRODUCTION

Thermal cameras measure long-wave infrared radiation ($\approx$ 8–14μm), capturing temperature distributions rather than reflected visible light [16]. Unlike RGB sensors, which record instantaneous intensity values in the visible range, thermal imaging provides access to heat transfer processes that often persist after an interaction has ended. These properties have made thermal imaging valuable in applications such as surveillance [2], medical monitoring [12], and environmental sensing [3].

Humans typically maintain a body temperature of about 37 °C / 98.6 °F, which is higher than that of most surrounding objects [9]. Physical contact with chairs, walls, or handheld items leaves behind residual heat patterns that decay over seconds to minutes, depending on the material properties and environmental conditions. These fading imprints function as passive temporal codes, providing not only information about the current scene but also indirect evidence of recent past events. While imperceptible in RGB imagery, such traces are readily observable in the thermal domain. As illustrated in Fig. 1, the residual heat marks intuitively indicate a prior interaction, giving observers a hint of "what just happened".

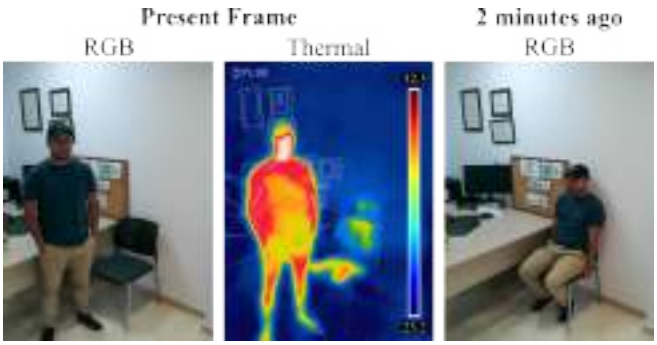

Fig. 1. Illustrative example of residual thermal traces. (***Left***) RGB and thermal images captured at the present frame. While the RGB view shows only the current visual appearance, the thermal image reveals heat imprints on the chair, acting as passive temporal codes and suggesting a prior human action. (***Right***) The ground-truth frame of the interaction that occurred approximately two minutes earlier.

Most prior work in temporal reasoning has concentrated on predicting the future from observed frames. RGB Video frame extrapolation methods such as Deep Voxel Flow [6] synthesize intermediate frames by estimating spatiotemporal voxel flows, while recurrent architectures such as PredNet [7] and CrevNet [18] propagate latent scene representations forward in time to forecast subsequent frames. Although designed for prediction, these models demonstrate that temporal structure can be learned and, if the conditioning were reversed, could in principle be adapted to infer plausible past sequences. However, inference methods based solely on RGB remain highly ambiguous, as many different past trajectories may be equally consistent with the same observed frames. To reduce this uncertainty, residual cues from other modalities have been explored. For example, in [15] an RGB–thermal dataset was introduced in which residual heat imprints en-

abled recovery of a person's pose up to three seconds earlier. While promising, this work is limited to pose-level inference, short temporal horizons, and does not address the reconstruction of full-scene appearance. To the best of our knowledge, no previous work has directly tackled the problem of reconstructing complete past scene states from thermal traces.

In this paper, a proof-of-concept for time-reversed scene reconstruction using paired thermal and RGB inputs is presented. The proposed framework combines Visual–Language Models (VLMs) with a constrained diffusion process: one VLM generates semantic descriptions of the scene conditioned on RGB and thermal inputs, while these descriptions guide the diffusion model to synthesize a plausible past frame. By exploiting faded thermal signatures as passive temporal codes and RGB as spatial context, the proposed method reconstructs visually consistent past images from present observations. Three representative cases are evaluated: sitting on a chair, leaning against a wall, and touching a book. Reconstructions are assessed using both low-level metrics (PSNR, SSIM) and high-level metrics (pose estimation, object detection). The results show that even in controlled scenarios, plausible recovery of past frames is achievable, marking the first step toward scene-level time-reversed imaging.

2. METHODOLOGY

2.1. Problem Formulation

The problem of time-reversed scene reconstruction is formulated as the task of inferring a plausible past image $\mathbf{x}_{RGB}^{t-\Delta}$ from current static multimodal observations. Specifically, access is assumed to an RGB frame $\mathbf{x}_{RGB}^{t} \in \mathbb{R}^{h\times w\times 3}$ and a co-registered thermal measurement $\mathbf{x}_{Thermal}^{t} \in \mathbb{R}^{h\times w}$ capturing residual heat traces left by human–environment interactions. The objective is to exploit these complementary modalities to recover a physically consistent estimate of the scene $\hat{\mathbf{x}}_{RGB}^{t-\Delta}$ at a short time horizon $\Delta > 0$.

2.2. Proposed Method

A framework is proposed for time-reversed scene reconstruction that leverages residual thermal traces, combined with RGB imagery for spatial context. The method consists of three main components: (i) multimodal input encoding, (ii) VLM guidance, and (iii) a constrained diffusion backbone for past-frame generation.

2.2.1. Multimodal Input Encoding

Given the current scene, a co-registered RGB image $\mathbf{x}_{RGB}^{t}$ and a thermal image $\mathbf{x}_{Thermal}^{t}$ are acquired at time $t = 0$. The RGB channel provides spatial detail, while the thermal modality encodes fading heat signatures caused by recent human–environment interactions. These two inputs are normalized and aligned before being passed to the VLM pipeline.

2.2.2. VLM-Based Guidance

The first step of the proposed method consists in producing a structured textual description of the current scene, conditioned on the multimodal inputs $(\mathbf{x}_{RGB}^{t}, \mathbf{x}_{Thermal}^{t})$ and the textual prompt p_{desc}. The resulting output p_{out} encodes high-level semantics (e.g., ``a person was recently sitting on the chair; residual heat is visible''), which serve as contextual cues to guide subsequent reconstruction. Importantly, we do not train a new model on RGB–thermal–text triplets. Instead, an existing pretrained VLM (GPT-5) is employed in inference mode, where the parameters θ^* remain frozen. This reduces the operation to a forward pass without gradient updates:

$$p_{out} = \arg\max_{y} \; p\big(y \mid \mathbf{x}_{RGB}^{t}, \mathbf{x}_{Thermal}^{t}, p_{desc}; \theta^*\big). \quad (1)$$

Here, p_{out} denotes the most likely textual description generated by the off-the-shelf VLM, given the RGB frame $\mathbf{x}_{RGB}^{t}$, the thermal frame $\mathbf{x}_{Thermal}^{t}$, and the conditioning prompt p_{desc}, under the frozen parameters θ^*.

2.2.3. Constrained Diffusion Framework

Then, the desired past image is generated using an existing pretrained diffusion model (Gemini 2.5 Flash Image, also known as NanoBanana), which was originally trained on large-scale general-purpose image datasets. In this method, the diffusion model is neither retrained nor finetuned; instead, its denoising trajectory is constrained by conditioning jointly on the multimodal inputs, the structured scene description p_{out}, and the generative prompt p_{gen}. This guidance ensures that the reconstruction remains semantically and physically consistent with both the observed modalities and the textual priors.

Formally, let $\mathbf{z}_T \sim \mathcal{N}(0, I)$ denote the initial Gaussian noise in the latent space. The pretrained diffusion model iteratively denoises $\mathbf{z}_t$ under multimodal constraints:

$$\mathbf{z}_{t-1} = g(\mathbf{z}_t, \mathbf{x}_{RGB}^{t}, \mathbf{x}_{Thermal}^{t}, p_{out}, p_{gen}; \theta^*), \quad t = T, \ldots, 1, \quad (2)$$

where g is the frozen denoiser guided through cross-attention mechanisms that integrate both visual and textual embeddings [13]. The final image is obtained via the pretrained decoder $f_{\text{dec}}(\cdot)$, which maps the latent representation back into RGB space:

$$\hat{\mathbf{x}}_{RGB}^{t-\Delta t} = f_{\text{dec}}(\mathbf{z}_0; \theta^*). \quad (3)$$

Equation 3 therefore represents the constrained generation of a plausible past scene $\hat{\mathbf{x}}_{RGB}^{t-\Delta t}$, obtained purely through inference with frozen parameters θ^*, where the prompts p_{out} and p_{gen} act as conditioning signals to guide the pretrained model toward reconstructions aligned with physically consistent semantics.

Fig. 2. Ablation study comparing RGB only, RGB+Thermal, and RGB+Thermal+Descriptor against ground truth, evaluated using low-level metrics (PSNR, SSIM) and high-level metrics (semantic segmentation with OA, pose estimation with MPJPE) to see the past, 30 seconds ago.

3. SIMULATIONS AND RESULTS

To validate the proposed method, a dataset was constructed comprising three controlled scenarios: *sitting on a chair*, *touching an object*, and *leaning against a wall*. In each case, a person maintained physical contact with the scene for 30 seconds, after which RGB and thermal images were acquired at multiple time delays ($5s, 15s, 30s, 120s$). Image acquisition was performed using a FLIR ONE camera, which provides aligned RGB and Long-Wave InfraRed (LWIR) thermal images. The RGB sensor has a resolution of 1440×1080, while the thermal sensor has a resolution of 80×60, operating at a frame rate of 8.7 Hz with a horizontal field of view of $55°$. The acquisition at $0s$ was taken as the ground-truth reference. Reconstructions of past frames were evaluated against the ground truth using both low-level and high-level metrics. For low-level metrics, the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) were implemented according to [8]. For high-level metrics, pose estimation with the Mean Per Joint Position Error (MPJPE) [15] and semantic segmentation with the Overall Accuracy (OA) [5] were employed.

Two pretrained VLMs were employed: GPT-5 for generating structured scene descriptions, and Gemini 2.5 Flash Image (NanoBana) as the conditional diffusion model for image generation. Fixed prompts were used for both tasks, as detailed below.

Prompt for Scene Description (p_{desc})

- `Identify the thermographic traces on the thermal image.`
- `In the RGB image, describe the scene.`
- `Output one concise past-tense sentence describing what happened some seconds or minutes ago, including all these details.`

Prompt for Generative Editing (p_{gen})

`Edit the RGB image using <description` p_{out}`> or <the thermal image> to depict the scene a few seconds earlier, preserving environment, lighting, angle, zoom, and colors.`

For the generative prompt, two options are considered in the ablation study, depending on whether the information from the scene description is included or not. Two main experiments were conducted: (i) an ablation study evaluating the contribution of the main components of the method. (ii) an assessment of how far into the past the proposed approach can reliably reconstruct in the controlled scenarios.

Table 1. Ablation study with different input modalities for the case of 30 seconds ago.

RGB	Thermal	Descriptor	High-level		Low-level	
			AO (↑)	MPJPE (↓)	PSNR (↑)	SSIM (↑)
✓	✗	✗	62.24 ± 0.75	336.92 ± 98.18	16.81 ± 0.54	0.76 ± 0.65
✓	✓	✗	67.55 ± 0.13	270.87 ± 13.02	17.22 ± 1.08	0.77 ± 0.14
✓	✓	✓	**76.87 ± 0.06**	**48.44 ± 14.18**	**18.57 ± 1.21**	**0.79 ± 0.02**

3.1. Ablation Study

To evaluate the impact of each input modality and the contribution of the scene descriptor within the overall framework, an ablation was conducted. Table 1 summarizes both high-level and low-level metrics under three configurations: (i) RGB input only, (ii) RGB combined with thermal input without the descriptor, and (iii) the full model incorporating RGB, thermal, and descriptor guidance. As illustrated in Fig. 2, the qualitative results align with the numerical trends. Using RGB-only inputs fails to reveal the previous state. Adding thermal data introduces complementary cues, allowing recovery of the person's prior action of sitting down a few seconds earlier. The most significant improvement, however, occurs when incorporating the scene descriptor: this semantic prior enables the model to infer higher-level interactions, such as detecting that the book was in hand. These results confirm that the VLM-based structured description acts as a strong prior, anchoring the diffusion process to physically consistent semantics.

Fig. 3. Input and ground-truth data for the supplementary material experiments. Each test scene includes the thermal image (left) capturing residual heat traces, the corresponding RGB image (middle), and the ground-truth action (right). Results are illustrated for Scene 1 (Test 1) and Scene 2 (Test 2)

3.2. Prompts descriptor Analysis

Figure 3 illustrates the input data used across the following two experiments. In each test, the model receives paired thermal and RGB images as input, where the thermal modality provides residual heat traces encoding the subject's recent interactions, and the RGB modality supplies contextual scene information. The ground-truth (GT) frame corresponds to the subject's actual position and action 20s earlier, serving as the reference for evaluation.

To study how instruction design affects "what-just-happened" descriptions from paired RGB–thermal inputs, we evaluate four descriptor prompts p^1_{desc}–p^4_{desc} (Fig. 4). Each prompt must output a single past-tense sentence stating the person's position and action 5 s earlier, while progressively adding explicit reasoning constraints: p^1_{desc} requests only the final sentence; p^2_{desc} requires enumerating primary/secondary heat traces in the thermal image; p^3_{desc} cross-checks those traces against the RGB image with object attributes and spatial relations; and p^4_{desc} adds a completeness check before synthesizing the final hypothesis. This progression isolates the effect of prompt structure, encouraging grounding on thermal residues and reducing ambiguity or hallucinations. Only the final sentence is returned; intermediate steps are used implicitly to guide inference. Empirically, more structured prompts yield descriptions that are more specific and consistent with the heat-trace evidence, especially in scenes involving multiple interactions (e.g., sitting while holding a book). The exact prompts used in the experiment are shown below, where the differences between the prompts and the previous one are highlighted in blue.

p^1_{desc} = *Follow strictly the step-by-step methodology below and do not skip or shorten any step before giving the final one-sentence answer.An RGB image and a thermal image of the same scene are attached.* ***Final output:*** *Provide only one short, direct sentence in past tense that precisely describes the person's position and action 20 seconds ago according to the heat traces. The answer must be concise and direct. Output only that sentence, nothing else.*

p^2_{desc} = *Follow strictly the step-by-step methodology below and do not skip or shorten any step before giving the final one-sentence answer. An RGB image and a thermal image of the same scene are attached. 1. Analyze the thermal image and list all objects or furniture that show any heat traces, starting with the one with the highest heat concentration (primary). Do not omit any object that shows heat, even if the trace is faint, small, or low intensity. Explicitly state when an object shows no heat. 2. Identify and describe all secondary heat traces (for example, on books, chairs, walls, desks, or other surfaces). Mention each one individually and describe its approximate intensity (strong, moderate, faint).* ***Final output:*** *Provide only one short, direct sentence in past tense that precisely describes the person's position and action 20 seconds ago, according to the heat traces. The answer must be*

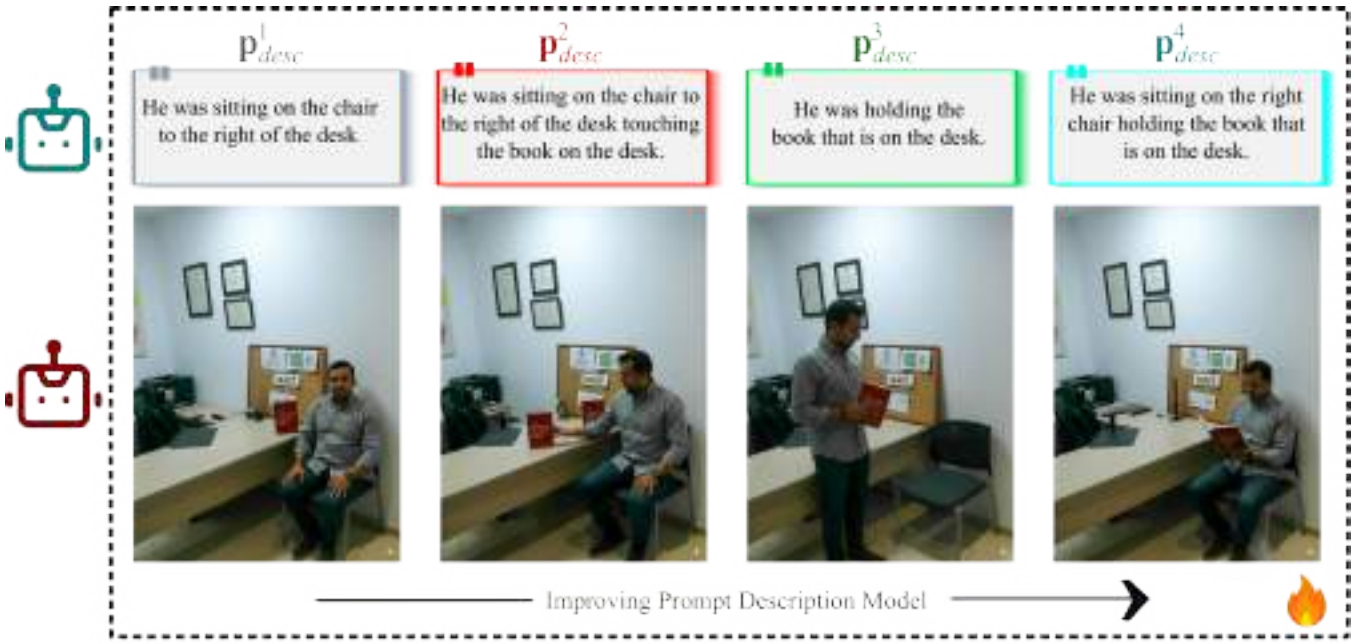

Fig. 4. Examples of past-action descriptions generated under different prompt descriptors p^1_{desc}–p^4_{desc}. The prompts progressively add reasoning constraints, from a simple direct description to structured analyses that incorporate thermal traces, RGB context, and completeness checks. More structured prompts yield descriptions that are more specific and consistent with the evidence.

concise and direct. Output only that sentence, nothing else.

p^3_{desc} = *Follow strictly the step-by-step methodology below and do not skip or shorten any step before giving the final one-sentence answer. An RGB image and a thermal image of the same scene are attached. 1. Analyze the thermal image and list all objects or furniture that show any heat traces, starting with the one with the highest heat concentration (primary). Do not omit any object that shows heat, even if the trace is faint, small, or low intensity. Explicitly state when an object shows no heat. 2. Identify and describe all secondary heat traces (for example, on books, chairs, walls, desks, or other surfaces). Mention each one individually and describe its approximate intensity (strong, moderate, faint). 3. Cross-check with the RGB image and locate every object with heat traces. For each object, provide: Object type, Object color, Position (left, center, right), Interaction with the person (touching, sitting, holding, near, none), Direction relative to the scene (front, back, left, right, corner) Final output: Provide only one short, direct sentence in past tense that precisely describes the person's position and action 20 seconds ago according to the heat traces. The answer must be concise and direct. Output only that sentence, nothing else.*

p^4_{desc} = *Follow strictly the step-by-step methodology below and do not skip or shorten any step before giving the final one-sentence answer. An RGB image and a thermal image of the same scene are attached. 1. Analyze the thermal image and list all objects or furniture that show any heat traces, starting with the one with the highest heat concentration (primary). Do not omit any object that shows heat, even if the trace is faint, small, or low intensity. Explicitly state when an object shows no heat. 2. Identify and describe all secondary heat traces (for example, on books, chairs, walls, desks, or other surfaces). Mention each one individually and describe its approximate intensity (strong, moderate, faint). 3. Cross-check with the RGB image and locate every object with heat traces. For each object, provide: Object type, Object color, Position (left, center, right), Interaction with the person (touching, sitting, holding, near, none), Direction relative to the scene (front, back, left, right, corner) 4. Before making the final inference, double-check that no object with visible heat traces has been left out of your analysis. 5. Based on both the primary and secondary heat traces, infer the most likely past action of the person 5 seconds ago. If multiple objects have heat traces, combine them into one concise sentence that mentions all relevant actions. Final output: Provide only one short, direct sentence in past tense that precisely describes the person's position and action 20 seconds ago according to the heat traces. The answer must be concise and direct. Output only that sentence, nothing else.*

3.3. Prompts Editor Analysis

To study how instruction design affects past-action *predictions* from paired RGB–thermal inputs, we evaluate four editing prompts p^1_{edit}–p^4_{edit} (Fig. 5). The editing model takes as input the current RGB image together with a prompt, and pro-

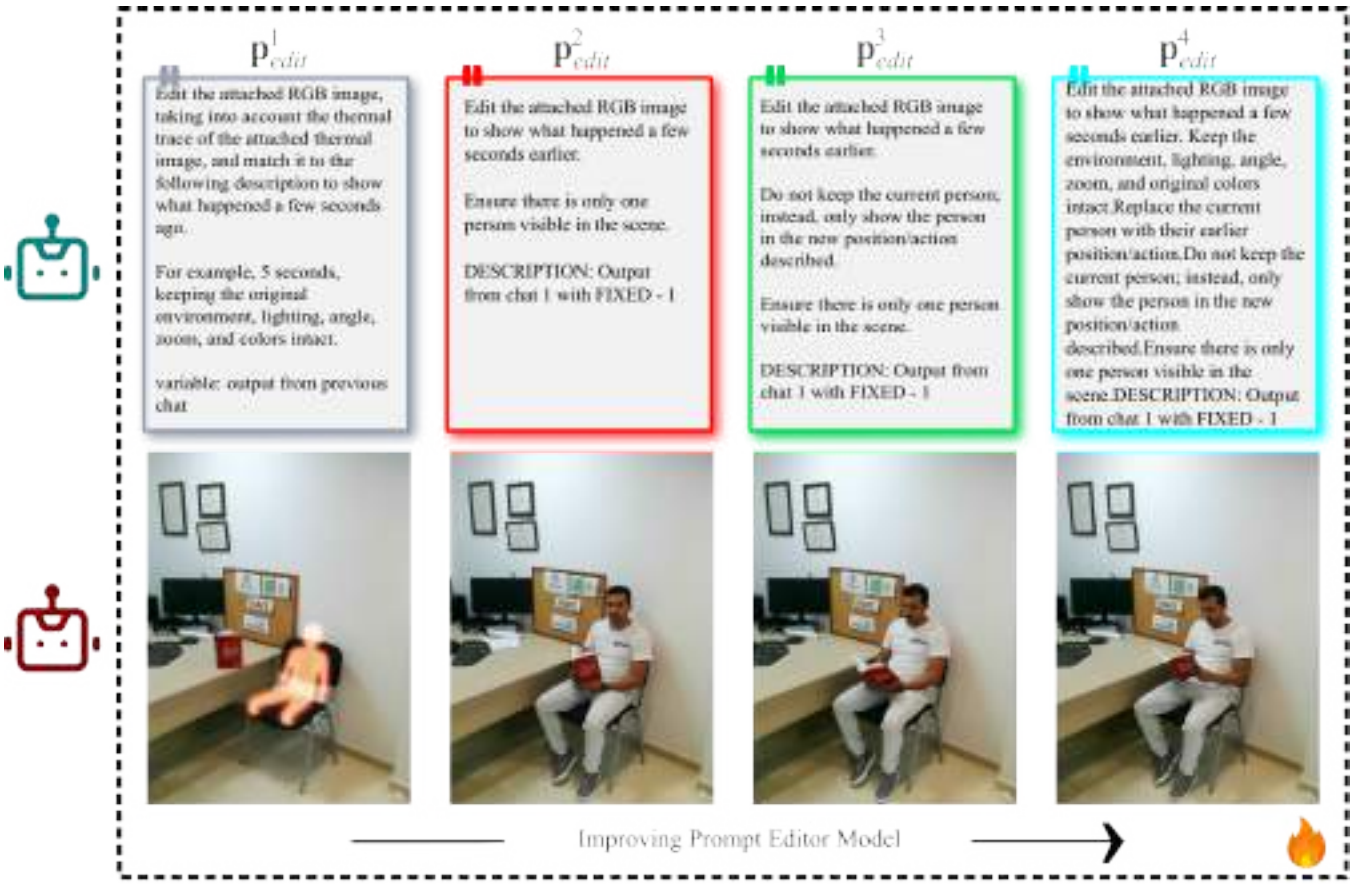

Fig. 5. Examples of past-action image predictions generated under different prompt descriptors p^1_{edit}–p^4_{edit}. Each prompt introduces progressively stricter editing constraints, from preserving the original environment with minimal changes to enforcing temporal consistency and position replacement. In all cases, the underlying action description used in the prompts was the same: *"the person was sitting and holding the book"*. The structured prompts guide the editing model toward outputs that are more coherent with the thermal trace, scene context, and past actions, demonstrating the impact of prompt design on editing quality and consistency.

duces an edited RGB output that predicts what happened a few seconds earlier. Each prompt progressively adds explicit constraints to guide the prediction: p^1_{edit} preserves the original scene setup (lighting, angle, and colors) while introducing temporal consistency with the thermal trace; p^2_{edit} enforces the presence of only one person in the scene; p^3_{edit} removes the current position and renders only the person in the new past position/action; and p^4_{edit} integrates both environment preservation and position replacement simultaneously. In all cases, the underlying action description is identical: *"the person was sitting and holding the book"*. This progression isolates the effect of prompt structure on the editing model, showing that more structured prompts yield past-action predictions that are more temporally consistent, visually coherent with the scene, and faithful to the intended description.

3.4. Image Generator Analysis

To investigate how different generative models perform in past-scene reconstruction, we employed the proposed scheme and only changed the generative model. For that we compared with DALLE 3 [10], Grok [17], Pixverse [11], Flux Kontext[1], Seeddream 4.0 [14], and Gemini 2.5 Flash Image [4]. All of them are VLM, where the generating process is constrained by the prompt used in our methodology. Figure 6 illustrates this comparison: the top row shows the reference RGB and thermal input image, while the following rows display the reconstruction produced by different generative models. This setup allows us to systematically assess how models vary in their ability to integrate multimodal cues, preserve environmental fidelity, and render temporally consistent human actions, thereby shedding light on their robustness in tasks that require reasoning about the immediate past.

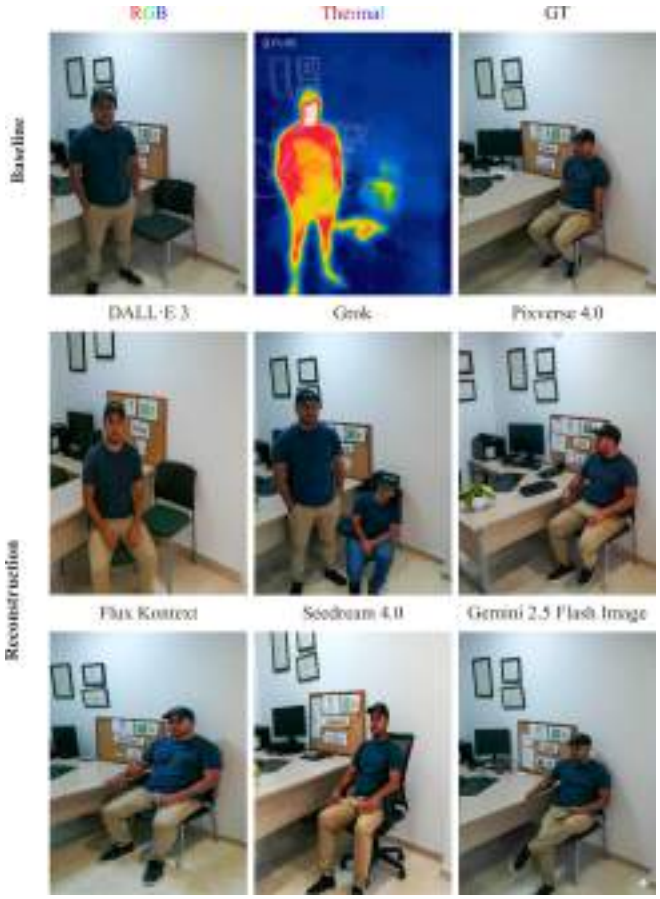

Fig. 6. Comparison between the image generators to estimate the past. The first row shows the inputs (RGB) and thermal, and the ground-truth (GT) scene. The subsequent rows illustrate past-scene reconstruction produced by different generative models.

3.5. Temporal Reconstruction Range

The last experiment assesses how far into the past the proposed method can reliably reconstruct a scene. Reconstructions were generated at increasing time delays ($5s, 15s, 30s$, and $120s$) and compared against the ground truth captured at $0s$. Performance was evaluated using both low-level and high-level metrics. As shown in Fig. 7, the top row depicts thermal frames, where the heat imprints progressively fade over time. The middle row presents the corresponding RGB reconstructions produced by the proposed framework, while the bottom row shows results from the RGB-only baseline. Notably, the RGB-only approach tends to hallucinate plausible but unconstrained frames, since it lacks the guidance provided by thermal information.

Figure 8 summarizes the quantitative results, indicating that the reconstructions remain stable up to $15s$, with only slight degradation in both fidelity and semantic accuracy. Beyond $30s$, however, the residual thermal traces weaken significantly, leading to noticeable declines in pose estimation

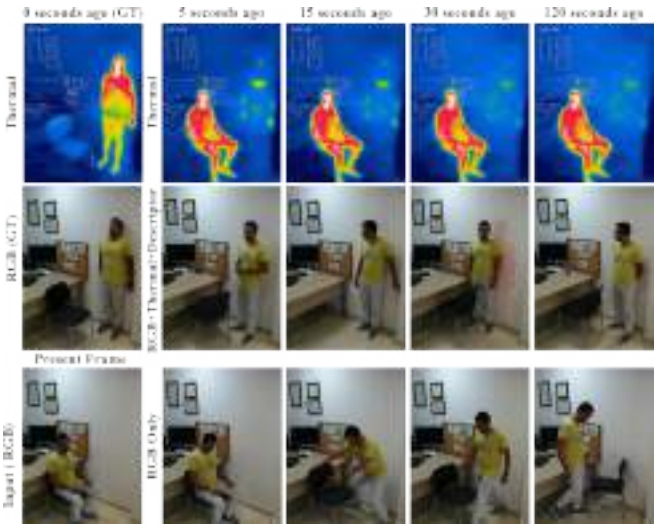

Fig. 7. Temporal reconstruction at increasing delays (0–120s). Top: thermal frames showing gradual heat decay. Bottom: RGB reconstructions guided by thermal traces and textual priors, recovering plausible past interactions up to two minutes after contact.

and overall accuracy. At $120s$, reconstructions still preserve the global structure of the scene but fail to capture fine details or precise interactions. These findings suggest that the effective temporal horizon of the proposed approach is between one and two minutes, after which the information available in residual traces becomes insufficient for reliable recovery.

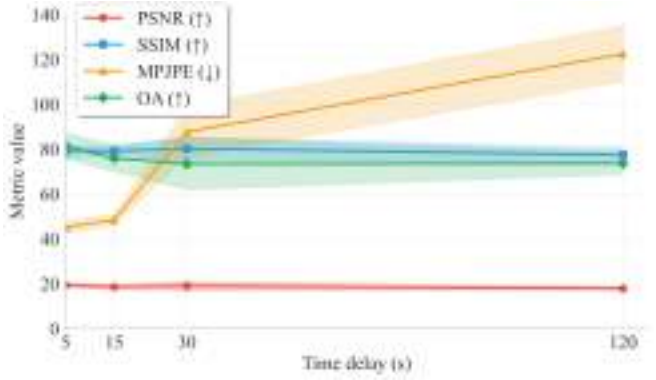

Fig. 8. Temporal reconstruction performance across increasing time delays. Both low-level metrics (PSNR, SSIM$\times$100) and high-level metrics (MPJPE, OA) are plotted.

4. CONCLUSIONS AND FUTURE WORK

This work presents a proof-of-concept framework for reconstructing recent past events by combining RGB and thermal imaging with VLM-guided diffusion models. To our knowledge, this is the first attempt to treat fading thermal imprints as temporal codes for scene reconstruction. Controlled experiments validate the feasibility of this approach, showing that

textual descriptions provide strong guidance and yield consistent gains across both low-level (PSNR, SSIM) and high-level (MPJPE, OA) metrics. The ablation study highlights the importance of VLM-based semantic priors, while temporal analysis demonstrates that past interactions can be plausibly recovered up to two minutes after contact. Future work will extend the framework to real-world environments with multiple subjects and variable illumination, and explore additional sensing modalities and efficiency optimizations for practical deployment in applications such as forensics, human–computer interaction, and security monitoring.

5. REFERENCES

[1] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025.

[2] James W Davis and Vinay Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer vision and image understanding*, 106(2-3):162–182, 2007.

[3] Bakhtiar Feizizadeh and Thomas Blaschke. Thermal remote sensing for land surface temperature monitoring: Maraqeh county, iran. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 2217–2220. IEEE, 2012.

[4] Google. Introducing gemini 2.5 flash image (aka nano banana). `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/`, 2025. Accessed: 2025-09-30.

[5] Carlos Hinojosa, Jorge Bacca, and Henry Arguello. Coded aperture design for compressive spectral subspace clustering. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1589–1600, 2018.

[6] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE international conference on computer vision*, pages 4463–4471, 2017.

[7] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[8] Brayan Monroy, Jorge Bacca, and Julián Tachella. Generalized recorrupted-to-recorrupted: Self-supervised learning beyond gaussian noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28155–28164, 2025.

[9] Daniel S Moran and Liran Mendal. Core temperature measurement: methods and current insights. *Sports medicine*, 32(14):879–885, 2002.

[10] OpenAI. Dall·e 3. `https://openai.com/dall-e-3`, 2023. Accessed: 2025-09-30.

[11] PixVerse. Pixverse ai video generator. `https://app.pixverse.ai`, 2025. Accessed: 2025-09-30.

[12] EFJ Ring and Kurt Ammer. Infrared thermal imaging in medicine. *Physiological measurement*, 33(3):R33, 2012.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[14] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

[15] Zitian Tang, Wenjie Ye, Wei-Chiu Ma, and Hang Zhao. What happened 3 seconds ago? inferring the past with thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17111–17120, 2023.

[16] Michael Vollmer and Klaus-Peter Möllmann. *Infrared thermal imaging: fundamentals, research and applications*. John Wiley & Sons, 2018.

[17] xAI. Grok. `https://x.ai`, 2025. Accessed: 2025-09-30.

[18] Wei Yu, Yichao Lu, Steve Easterbrook, and Sanja Fidler. Crevnet: Conditionally reversible video prediction. *arXiv preprint arXiv:1910.11577*, 2019.