Sudo Thermal Camera: Heat-Adjustable Generative Video From a Single RGB Image

Anonymous CVPR submission

Paper ID *****

Abstract

We present Sudo Thermal Camera, a training-free system that takes a single RGB image and a user-specified heat-source map as input to generate a physically plausible video showing what the scene would look like if that heat were applied. The system combines four non-learned components: (i) text-prompted material segmentation with SAM3, (ii) a 2D pixel-space thermal simulator implementing Fourier heat diffusion with phase change and reaction progress, (iii) heat-aware Cross-Attention Guidance (CAG) that perturbs cross-frame attention at three temporal-matching layers of a frozen HunyuanVideo-1.5 image-to-video diffusion transformer, localised by the simulator's time-evolving mask, and (iv) pixel-composite preservation locking non-heated pixels bit-equal to the input. Our model gates the CAG mask and pixel composite on SAM3 material labels, closing an identity-drift failure mode observed when the heat mask leaks onto non-responsive materials. We validate on six scenes spanning liquid/steam, wood combustion, metal glow, wax ignition, paper burn, and ice melt. The first five produce scene-appropriate motion; the ice case reveals a structural limitation of any inference-time attention-steering approach over a frozen prior.

1. Introduction

Generative video models such as HunyuanVideo [1], Sora, and CogVideoX [2] can synthesise long, realistic clips from a still image, but they have no notion of physics. Asking HunyuanVideo for "a kettle on a hot stove" produces a video in which the kettle may or may not steam, and even when it does, the steam has no causal relationship with how much heat was applied, where, or for how long.

We are interested in the inverse. Given a single RGB image and an explicit heat input (a 2D heat-source map painted by the user or measured by an IR camera), produce a video in which thermal effects evolve according to known physics. The output video should preserve the input image outside thecheated region and animate physically correct effects inside the heated region (boiling, melting, charring, glowing) appropriate to each material, and respect heat conduction so the effected region grows over time. Our contributions are as follows:

1. We generate the pipeline that decouples physics simulation from motion synthesis: a deterministic 2D Fourier-diffusion PDE with phase change and reaction progress handles thermal evolution; a frozen pretrained video model supplies visual realism localised to the simulator's heated region.
2. We propose Heat-aware Cross-Attention Guidance (Section 2.3): a training-free inference-time intervention that restricts CAG's attention perturbation to the intersection of the physics-derived heat mask and a SAM3 whitelist of thermally responsive materials. No learned parameters and the entire backbone is frozen.
3. A pixel-composite preservation mechanism that locks non-heated pixels bit-equal to the anchor image, with per-material anchor retention in V2 for scenes where non-responsive materials share a mask with responsive ones (Section 2.4).
4. We generate a FLIR ONE-based dataset of 7 indoor heating scenes (573 paired pre-heat/heated frames); we used for IR-derived heat maps and true heated endpoints in evaluation.

2. Method

Figure 1 shows the inference pipeline. SAM3 segments the anchor into materials. Then the user's per-material heat index ΔQ is rasterised into a continuous 2D heat-source map $Q_{\text{ext}} \in [0,1]^{H\times W}$. The PDE produces a per-frame (T, φ, r) stack. The physics-derived mask intersected with the SAM3 heated-material whitelist gates cross-attention perturbation at three temporal-matching DiT layers. Eventually, a pixel-composite with per-material anchor retention produces the final video.

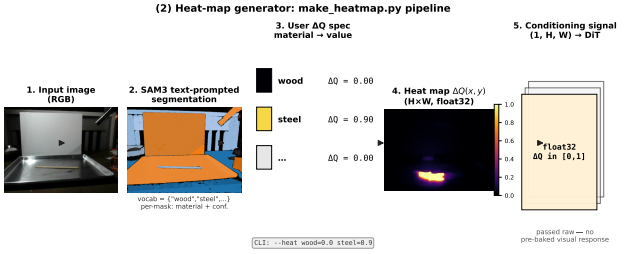

Figure 1. Heat-map generator and conditioning signal. SAM3 labels each pixel with a material; the user specifies a scalar heat index $\Delta Q \in [0, 1]$ per material; the rasterised float32 map is passed raw to both the 2D PDE and the downstream DiT.

2.1. Material segmentation (SAM3)

We use Meta's pre-trained frozen SAM3 [3] to label every pixel with one of 12 supported materials.

$$\mathcal{V} = \{\text{air, ice, water, wax, wood, wick, steel, copper, plastic, glass, stone, unknown}\} \quad (1)$$

SAM3 returns segments $\{(M_i, m_i, c_i)\}$. Here segments with $c_i < 0.2$ or area < 1000 pixels are dropped. Overlaps resolve by painting smaller masks on top. The per-pixel material map feeds three downstream consumers which is PDE property lookup, CAG mask gating, and per-material anchor retention.

Each pixel is a $\Delta x = 1$ mm cell; we integrate by explicit Euler with CFL-safe sub-stepping. Five per-material fields are looked up from Table 1. For each material name we look up five physical properties from a fixed table (Table 1):

- $\kappa(x, y)$ — thermal conductivity $[\mathrm{W}/(\mathrm{m} \cdot \mathrm{K})]$
- $c_p(x, y)$ — specific heat capacity $[\mathrm{J}/(\mathrm{kg} \cdot \mathrm{K})]$
- $T_{\text{phase}}(x, y)$ — phase-transition temperature [K]
- $r_{\text{rate}}(x, y)$ — reaction (combustion) rate constant $[1/\mathrm{s}]$
- $T_0(x, y)$ — initial temperature [K]

Density is fixed at $\rho = 10^3$ kg/m^3.

2.2. Physics-Based Thermal Simulation

2.2.1. Heat Diffusion Equation

Thermal evolution is governed by Fourier's law augmented with an external volumetric source:

$$\frac{\partial T}{\partial t} = \alpha(x, y)\, \nabla^2 T + \frac{Q_{\text{ext}}(x, y)}{\rho\, c_p(x, y)}, \qquad \alpha(x, y) = \frac{\kappa(x, y)}{\rho\, c_p(x, y)}, \quad (2)$$

where ∇^2 denotes the discrete 5-point Laplacian applied with replicate boundary conditions. The external heat source is defined as $Q_{\text{ext}} = s_h \cdot \sigma_Q \cdot Q$, where $Q \in [0, 1]^{H \times W}$ is the spatial heat map. The baseline calibration scale is empirically set to $\sigma_Q = 2 \times 10^8$ W/m^3, establishing a condition where maximum heat application ($Q = 1$) on water induces a phase transition within a standard 30-frame simulation interval at $\Delta t = 0.1$ s. An adjustable multiplier s_h allows for direct modulation of the overall heat intensity.

2.2.2. Phase Transition Modeling

A latent-heat-aware progress variable $\varphi(x, y, t) \in [0, 1]$ tracks the fraction of the local material that has phase-changed (e.g., melted, vaporised):

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t} = \frac{c_p\, \max\big(0,\, T - T_{\text{phase}}\big)}{L_{\text{latent}}}. \quad (3)$$

While $\varphi < 1$, the temperature is clamped at T_{phase} to absorb excess thermal energy into latent heat:

$$T \leftarrow \begin{cases} T_{\text{phase}}, & \text{if } \varphi < 1 \text{ and } T > T_{\text{phase}}, \\ T, & \text{otherwise.} \end{cases} \quad (4)$$

We use $L_{\text{latent}} = 3.34 \times 10^5$ J/kg (water's latent heat of fusion) as a single-material approximation.

2.2.3. Combustion and Reaction Progress

For materials with non-zero reaction rate (wax, wood, wick, plastic), we track an irreversible burn fraction $r(x, y, t) \in [0, 1]$:

$$\frac{\mathrm{d}r}{\mathrm{d}t} = r_{\text{rate}}\, \max\big(0,\, T - T_{\text{phase}}\big). \quad (5)$$

2.2.4. Numerical Integration and Stability

We use explicit Euler with sub-stepping to satisfy a 2D-Laplacian CFL constraint:

$$\Delta t_{\text{safe}} = 0.2 \cdot \frac{\Delta x^2}{\max_{x,y} \alpha(x, y)}, \qquad n_{\text{sub}} = \lceil \Delta t / \Delta t_{\text{safe}} \rceil, \qquad \delta t = \Delta t / n_{\text{sub}}. \quad (6)$$

For every video frame the solver advances Eqs. 2, 3, and 5 for n_{sub} inner steps of size δt, then snapshots (T, φ, r) into the per-frame physics stack.

2.3. Heat-aware Cross-Attention Guidance

We locate HunyuanVideo-1.5-I2V's temporal-matching layers with DiffTrack [5] and find that three mid-depth blocks $\mathcal{L}^{\text{CAG}} = \{16, 21, 25\}$ account for most cross-frame confidence $C^{\text{cross}}_{l,t}$. Standard Cross-Attention Guidance [4] perturbs cross-frame attention globally. We restrict the perturbation to the intersection of the PDE-derived heat mask and a SAM3 heated-material whitelist $\mathcal{H}$:

$$M^{\text{V1}}_{\text{CAG}}(t, p) = \mathbb{I}[h(t, p) \geq \tau], \quad (7)$$

$$M^{\text{V2}}_{\text{CAG}}(t, p) = M^{\text{V1}}_{\text{CAG}}(t, p) \wedge \mathbb{I}[s(p) \in \mathcal{H}], \quad (8)$$

where $h(t, p) = \mathrm{clip}(\Delta T / T_{\text{span}} + \varphi + r, 0, 1)$ is the physics-derived heat field, $\tau = 0.5$, and $\mathcal{H}$ is the SAM3 heated-material whitelist:

$$\mathcal{H} = \{\text{water, wax, wood, wick, steel, copper, ice, glass, plastic}\}. \quad (9)$$

Table 1. Per-material thermal properties (subset) used by the 2D PDE. $\rho = 10^3$ kg/m^3 fixed.

Material	κ	c_p	T_{phase}	r_{rate}	T_0
water	0.6	4182	373	0.0	293
wax	0.25	2950	330	0.1	293
wood	0.17	1700	573	0.5	293
wick	0.10	1300	523	2.0	293
steel	50.0	500	1811	0.0	293
ice	2.2	2090	273	0.0	263
copper	385	385	1358	0.0	293
glass	1.0	840	1700	0.0	293
plastic	0.20	1400	430	0.3	293

Cross-frame attention entries between two heated tokens are set to $-\infty$ on the degraded branch and the guided prediction $\tilde{\epsilon} = \epsilon_{\text{u}} + s_{\text{cfg}}(\epsilon_{\text{c}} - \epsilon_{\text{u}}) + s_{\text{cag}}(\epsilon_{\text{c}} - \hat{\epsilon})$ is used by the scheduler. The whole mechanism adds no trainable parameters.

2.4. Pixel-composite preservation

To stop DiT self-attention leakage from modifying cold pixels, we composite the generated video against the anchor at output resolution:

$$V_{\text{out}}(t) = M_{\text{pix}}(t)(1 - a(p))\, V_{\text{gen}}(t) + (1 - M_{\text{pix}}(t)(1 - a(p)))\, I_{\text{anchor}} \quad (10)$$

where $M_{\text{pix}}(t)$ is the per-frame preserve mask and $a(p) = a_{s(p)}$ is a per-material anchor retention looked up from a small empirical table ($a_{\text{ice}} = 0.8$, $a_{\text{wax}} = 0.7$, $a_{\text{water}} = 0.2$, $a_{\text{steel}} = 0.1$, …). V1 uses a global scalar $a^\star$; V2 replaces it with the per-material map. Pixel-space composite avoids the VAE round-trip artefacts that plagued our earlier latent-level attempts (RePaint blending, hidden-state freezing, flow-matching velocity substitution).

3. FLIR ONE Dataset

Figure 2 summarises the FLIR ONE dataset. Each row shows an anchor RGB (pre-heat), the final heated RGB (peak), the raw MSX infrared frame, and the calibrated thermal heat map after cleaning; scenes span paper ignition, wood char, steel glow, ice melt, water boil, and candle wick combustion.

3.1. Capture

We collected 1712 dual-resolution MSX JPEGs with a FLIR ONE Pro Gen 3 (640×480 thermal sensor + 1440×1080 visible camera). Each scene captures a deliberate heating event: *tissue_1, tissue_2* (paper ignition with a lighter), *chopstick_1, chopstick_2* (wood char), *fork_heated* (steel thermal glow), *ice_torch* (ice melting), *water_kettle* (water boiling on stove), and four candle scenes (omitted from training, see Section 3.4). For the capturing protocol, we used tripod- mounted FLIR ONE, scene set up and recorded for ∼10s at ∼1 Hz. Each scene contributes a contiguous range of frame IDs (e.g., `IMG_4525–IMG_4585` for *tissue_1*).

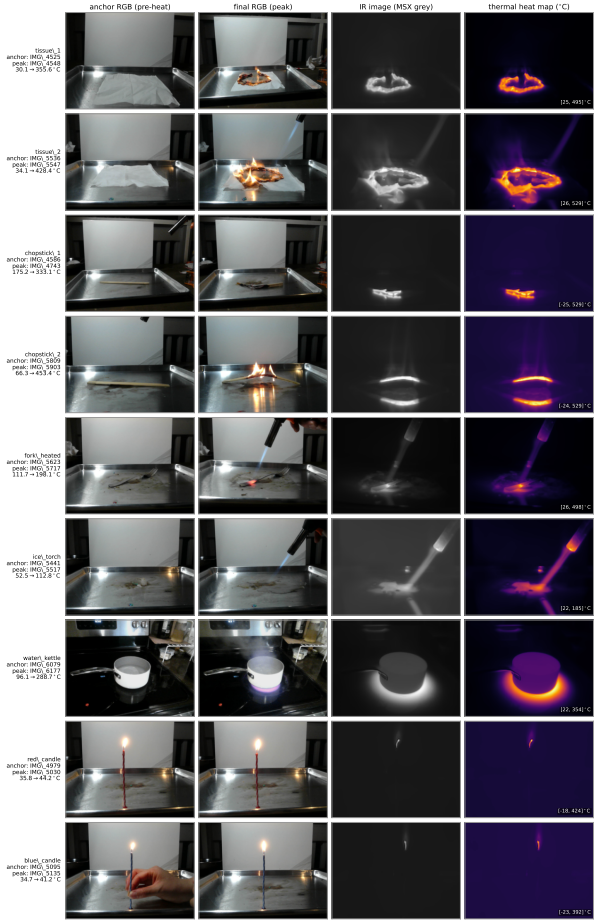

Figure 2. FLIR ONE Pro dataset. Rows span the scenes listed in Section 3.1: *tissue_1/2* (paper ignition), *chopstick_1/2* (wood char), *fork_heated* (steel glow), *ice_torch* (ice melt under torch), *water_kettle* (water boil), and the four candle scenes (filtered out during training, see Section 3.4). Columns are anchor RGB (pre-heat frame used as I2V conditioning), final RGB (peak of the heating window), extracted IR intensity image, and the calibrated per-pixel thermal heat map.

3.2. Preprocessing

The raw thermal blob embedded in each MSX JPEG is extracted with `flyr` and aligned to the visible image. The FLIR ONE Gen 3 sensor exhibits three failure modes that we clean before training:

1. **Dead pixels (NaN).** Hardware-fixed pixels report NaN regardless of scene content. Across the dataset, 734 frames contain NaN values; all affected frames have

$\leq 1\%$ NaN pixels concentrated in short rows or columns (consistent with the sensor's known dead-row pattern). We replace NaN values with the nearest finite neighbour using a single 2D Euclidean distance transform (`scipy.ndimage.distance_transform_edt` with `return_indices=True`). This is deterministic, $\mathcal{O}(N)$, and preserves material boundaries because we copy a single measured pixel rather than averaging.

2. **Saturation (Inf).** The sensor's measurement range is -20 °C to 400 °C; values beyond are reported as $\pm\infty$. We clamp $+\infty \rightarrow 400$ °C and $-\infty \rightarrow -20$ °C, preserving the "saturated hot/cold" information rather than discarding.
3. **Calibration glitches.** Adjacent extreme temperature gradients (e.g., a 500 °C flame next to a 20 °C wall) can break the sensor's per-row calibration and produce isolated finite values like -131 °C in an indoor scene. We treat values outside the plausibility range $[-25, 600]$ °C as bad and apply the same nearest-neighbour fill.

Frames with $> 5\%$ bad pixels are dropped as genuinely corrupted. After cleaning, we recover 1711 of 1712 raw frames (a $+75\%$ gain over the naive NaN-drop baseline of 977 frames).

3.3. Pair manifest construction

For each scene we automatically detect three temporal phases by analysing the 99$^{\text{th}}$ percentile temperature time-series of `thermal_raw/*.npy`:

- pre-heat — frames before the temperature exceeds baseline $+ 0.10 \cdot$ (peak $-$ baseline),
- heating — from the first heated frame up to the peak,
- cool-down — from the peak to the last heated frame.

Manual inspection revealed that for several scenes (e.g., *chopstick_2*), the last pre-heat frame already had the lighter visible at > 250 °C, contaminating the "cold baseline" assumption. We therefore use the first pre-heat frame as the anchor. `water_kettle` is the only exception, where the only available pre-heat frame already shows the hot stove element (a valid scene-level baseline).

A training pair is $(\mathbf{I}^{\text{anchor}}, \mathbf{I}^{\text{target}})$ where the target is any frame from the heating range plus the first 25% of the cool-down range. The dataset yields 573 pairs (Table 2).

3.4. Filtered scenes

The four candle scenes (*green_candle_1, green_candle_2, red_candle, blue_candle*) all have peak-minus-baseline temperature spans < 13 °C at the 99$^{\text{th}}$ percentile because the wick flame occupies $< 0.2\%$ of the image and is diluted by the cold wax background. They are excluded from training.

Table 2. Training pair counts and thermal characteristics per scene.

Scene	Anchor frame	Pairs	Peak T [°C]	ΔT [°C]
tissue_1	IMG_4525	21	356	326
tissue_2	IMG_5536	14	428	394
chopstick_1	IMG_4586	160	333	158
chopstick_2	IMG_5809	110	453	387
fork_heated	IMG_5623	91	198	86
ice_torch	IMG_5441	70	113	60
water_kettle	IMG_6079	107	289	193
Total		**573**		

4. Experiments

4.1. V1 results at scene scale

Three scenes covering physically distinct categories (*water_kettle*, *chopstick_1*, *fork_heated*), identical config ($s_{\text{cag}} = 1.0$, window $[30, 45)$, blank prompt, adapter-free):

Scene (category)	bg_MSE↓	hot_MSE↓	hot_temp
water_kettle: std CAG (MVP)	798.2	5640.2	4.49
water_kettle: V1 heat-aware	**322.6**	**2716.9**	3.32
chopstick_1: MVP	736.3	1458.9	5.46
chopstick_1: V1	**92.3**	**1377.0**	2.13
fork_heated: MVP	667.4	3779.1	14.45
fork_heated: V1	**558.3**	**3031.6**	6.90

V1 reduces bg_MSE by 16%–88% over the unmasked-CAG MVP with simultaneous reductions in hot_MSE and temporal distance. All three scenes produce category-appropriate motion (steam plume, flame-like char glow, metal-glow tint) purely from HunyuanVideo's prior under heat-localised attention steering.

4.2. V2 SAM3-gated ablation on water_kettle

V1 produces an identity-drift artefact (F5, a ghost blue person figure at seed 42) when the heat mask leaks onto non-responsive materials. V2 closes this by (i) intersecting the CAG mask with the SAM3 heated-material whitelist (B1), and (ii) replacing the scalar $a^\star$ with a per-material map indexed by SAM3 label (B2). A third variant (B_clean) uses B1+B2's pixel gate but keeps a small scalar $a^\star$ to tune hot-region generation freedom. Table 3 reports end-vs-start diff-energy (lower = less change) with annotated qualitative outcomes.

4.3. Out-of-distribution limits

We additionally tested three OOD categories: wax combustion, paper tissue burn, and ice melting. Wax succeeds; paper burn succeeds modulo primitive placement; ice fails semantically. This is becuase HunyuanVideo's prior for "vis-

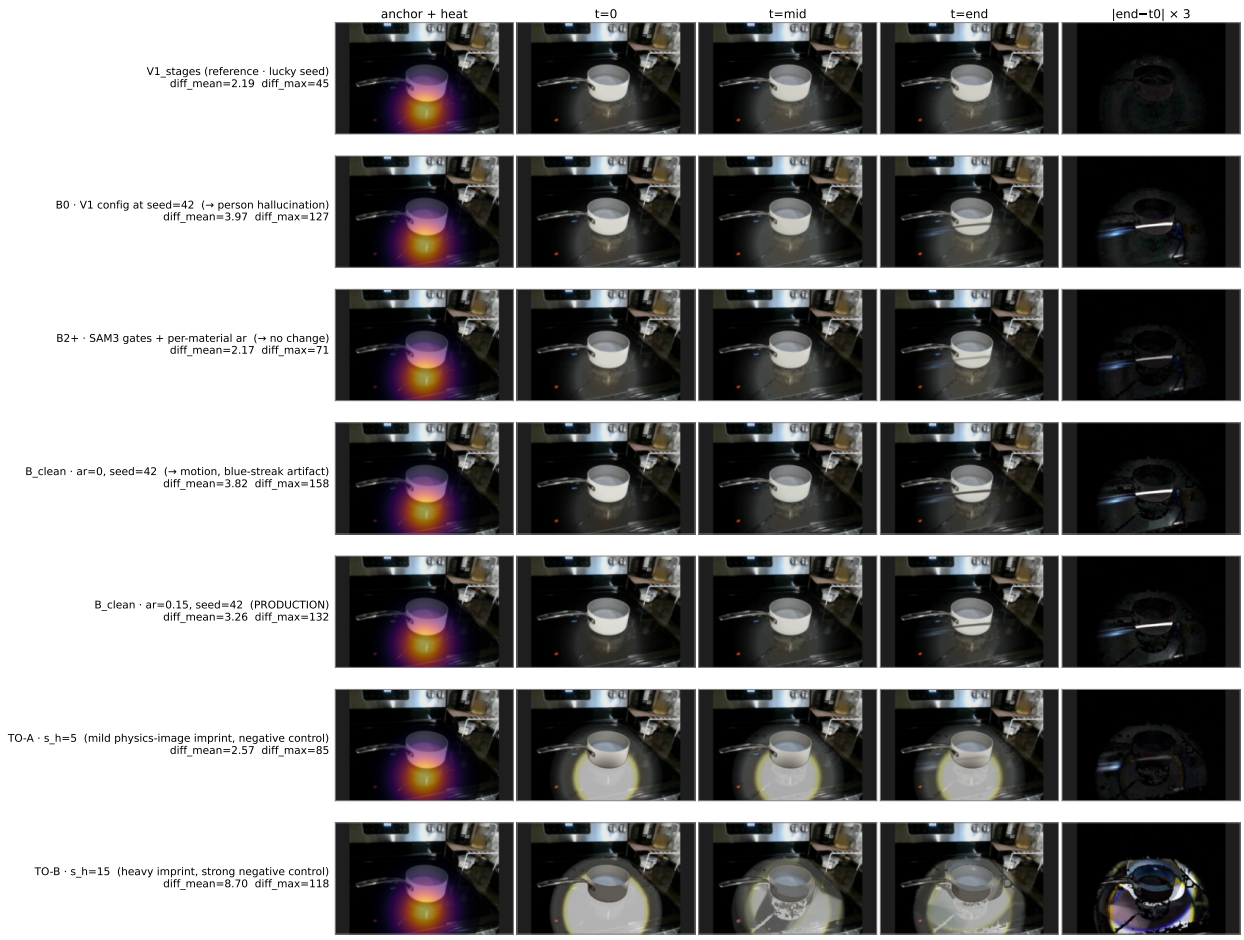

Figure 3. V2 1 s ablation panel on *water_kettle*. Rows top-to-bottom match Table 3. Columns: anchor+heat overlay, $t = 0$, $t = \text{mid}$, $t = \text{end}$, $|t_{\text{end}} - t_0| \times 3$ diff. B_clean at $a^\star = 0.15$ (production) shows a clean kettle-rim shimmer with no ghost-person drift. The s_h sweep (TO-A/TO-B) is a negative control: raising the physics forcing just stamps the physics image onto the anchor rather than producing a plausible thermal response.

Table 3. V2 1 s ablation on *water_kettle*, seed 42, thermal-only.

Config	diff_mean	diff_max	Behaviour
V1_stages (lucky seed)	2.19	45	subtle shimmer
B0 ($a^\star = 0$)	3.97	127	blue-person (F5)
B2+ per-material a_m	2.17	71	clean, no change
B_clean ($a^\star = 0$)	3.82	158	blue streak
B_clean ($a^\star = 0.15$)	**3.26**	**132**	**clean, shimmer**
TO-A ($s_h = 5$)	2.57	85	physics imprint
TO-B ($s_h = 15$)	8.70	118	heavy imprint

ible heat source → flame" outweighs its prior for "ice melting under applied heat", and CAG amplifies the stronger prior. This is a structural limitation of any inference-time attention-steering approach over a frozen prior. Our model localize and amplify existing priors but do not manufacture new ones.

5. Discussion

What's deterministic vs. learned. Heat conduction, phase change, and reaction progress are deterministic outputs of Eqs. 2–10 and are not invented by the network. The visual translation into bubbles, steam, char, and glow comes from HunyuanVideo's pretrained prior, localized and amplified by heat-aware CAG. No model parameters are trained and the text prompt is held blank so all semantic steering comes from the thermal signal alone.

Why 2D suffices. The surface-level effects we need (boiling, melting, charring, glowing) are all visible at the image plane. A 2D PDE is ∼100× cheaper than a 3D MPM/Triplane stack and adequate at our target resolutions.

References

[1] Hunyuan Video Team. HunyuanVideo-1.5: a publicly available 8.3B-parameter text-and-image-to-video diffusion transformer, 2025. 1

[2] Yang, Zhuoyi et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 1

[3] Meta FAIR. Segment Anything Model 3 (SAM3), 2025. 2

[4] Hong, Sungmin et al. Cross-Attention Guidance for sampling-time control of diffusion models, 2024. 2

[5] DiffTrack authors. DiffTrack: Locating temporal-matching layers in video diffusion transformers, 2025. 2

[6] Zhang, Lvmin et al. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023.

[7] Geng, Daniel et al. Motion Prompting: controlling video generation with motion trajectories. In *CVPR*, 2025.

[8] [Force Prompting authors]. Force Prompting: physical-control video generation, 2025.

[9] [PhysDreamer authors]. PhysDreamer: physics-based interaction with 3D objects via video generation, 2024.

[10] HDRT Dataset, 2024.

[11] Peng, Bohao et al. ControlNeXt: Powerful and efficient control for image and video generation. arXiv:2408.06070, 2024.

[12] Tan, Zhenxiong et al. OminiControl: Minimal and universal control for diffusion transformer. arXiv:2411.15098, 2024.

[13] Lugmayr, Andreas et al. RePaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

[14] Couairon, Guillaume et al. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *ICLR*, 2023.

[15] Cao, Mingdeng et al. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023.

[16] Rout, Litu et al. Semantic image inversion and editing using rectified stochastic differential equations. arXiv:2410.10792, 2024.

[17] Kulikov, Vladimir et al. FlowEdit: Inversion-free text-based editing using pre-trained flow models. arXiv:2412.08629, 2024.