VECTILE: LEARNING REUSABLE TRANSPORT REFERENCES FOR COUNTERFACTUAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A visual reference shows a response under particular conditions; editing often requires that response under another stimulus, installation, or arrangement of interacting objects. We introduce transport references: reusable local responses rather than the trajectories or illumination patterns observed in a reference. VECTILE extracts a response tile, installs it at a target site, and rolls out the installed tiles through TILEFLOW, where the target repeatedly supplies their current inputs and states. Visual synthesis then expresses the resulting consequence in images or videos. We show that this separation supports both optical interactions, where outgoing influence changes what other sites receive, and active-material deformation, where changing shape alters subsequent exposure. Our optical implementation couples local response and typed transport to spatial-goal-guided diffusion; the material implementation supplies camera-specific controls from one shared predicted world trajectory. Across 68 configurations of a held-out active-material world, VECTILE produces 52 correct edits versus 6 for our VIPER-style/Wan implementation, with 16 low-visibility configurations. On a relighting benchmark, VECTILE recovers seven edits missed by our natively adopted LightMover-style/Wan baseline.

1 INTRODUCTION

Reference-guided generation has progressed beyond appearance matching. LightMover controls emitter position, color, intensity, and illumination (Zhou et al., 2026); VIPER uses reference video to condition physical behavior (Chen et al., 2026). This raises a question about the unit of transfer: what should remain the same when the target calls for a different result? A reference sheet bends under a particular stimulus and mounting. Reinstalled along another axis, it should deform in another direction; starting warm, it may keep bending after heating stops. The observation contains both a reusable response and conditions specific to the reference scene.

Light editing presents the same separation. Moving an emitter should preserve the responses of unedited surfaces, but not their old highlights: incoming light and downstream illumination must follow the target. Interaction makes the distinction sharper. Replacing a front sheet's response can expose or shield a rear sheet. The rear keeps its reference, yet its motion should change. The desired operation is a *local replacement with coupled consequences*, not the replay of a trajectory or a fixed whole-scene edit.

Our hypothesis is that a reference should preserve a *way of responding*, while the target determines what happens next. We call this semantic unit a *response tile*: a reference-derived code evaluated through a shared local-response model. Installation assigns a site and frame, but the target supplies the current input and state. Reuse therefore preserves the local transformation, not its output under the reference conditions. For a specified form of conditional composition, Bradley (2026) relates compositional structure to sparse score dependencies, including feature-space dependencies. Our choice is to localize reference dependence at an interaction site, not to limit the spatial reach of its consequences.

This choice leads to four stages: **Extract** → **Install** → **TileFlow** → **Visual Synthesis** (Fig. 1). TILEFLOW repeatedly evaluates installed tiles and uses their outputs to update the target inputs for subsequent calls. Radiative transport suggests this alternation of local response and target-dependent propagation (Pharr et al., 2023). In active materials, the corresponding feedback passes through state and shape, then renewed exposure. We show that local responses can remain reusable as their

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

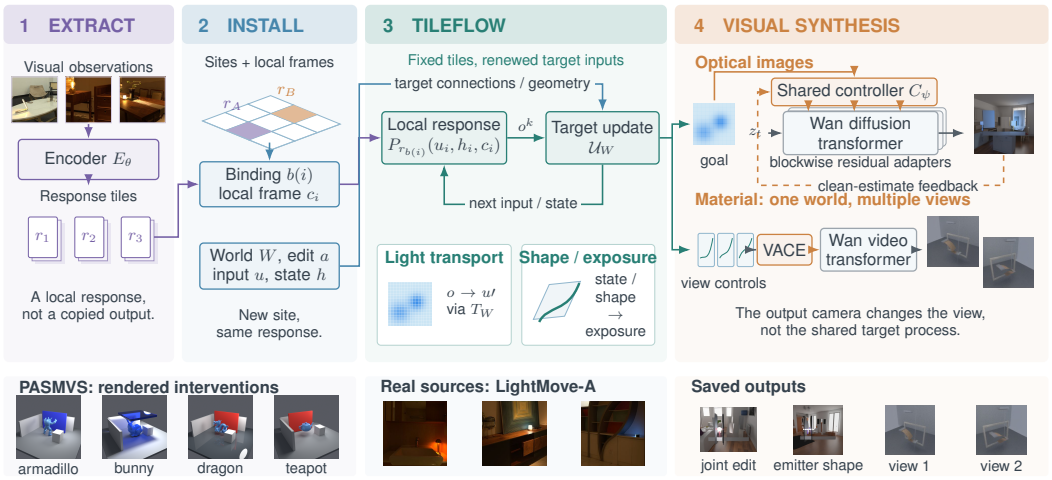

Figure 1: Extract → Install → TileFlow → Visual Synthesis. A fixed response tile is called on inputs renewed by the target. Optical synthesis uses a spatial goal and feedback; material synthesis observes one predicted world from multiple cameras. The lower strip shows how the data support the study: PASMVS-based rendered pairs (Broekman and Gräbe, 2020) provide controlled material interventions, LightMove-A supplies real-image editing contexts, and generated examples show optical edits and multi-view material responses.

consequences pass through optical transport or through changing shape and exposure. The material tests examine reuse and replacement; the optical tests examine how target-dependent consequences become visible.

A computed consequence is not yet a generated image: a receiver represented in a transport field may remain unchanged after synthesis. We distinguish TILEFLOW from *Visual Synthesis*. Optical interaction features predict a spatial goal that repeatedly guides diffusion. Material dynamics produce one world trajectory, projected into video controls for each camera. Re-observation changes projection and appearance, not response identity or interaction history. For interactive scene editing and generative world modeling, this lets users inspect or revisit an intervention from different viewpoints while retaining one underlying process.

Our contributions are:

1. We introduce *transport references*: independently installable response tiles that preserve a way of responding while allowing its target-dependent consequences to change.
2. We build VECTILE around *Extract, Install, TileFlow, and Visual Synthesis*, coupling reusable local responses to recurrent target updates and domain-specific diffusion interfaces for optical transport and active-material deformation.
3. We demonstrate response reuse across state, installation, and observation changes; composition in unseen ordered pairings and joint edits; and improved optical receiver activation and emitter placement.

2 RELATED WORK

Reference-guided visual editing. LightMover controls emitter position, color, and intensity; VIPER conditions generation on demonstrated physical behavior (Zhou et al., 2026; Chen et al., 2026). They are our direct named comparisons. ControlNet and IP-Adapter add spatial and image conditioning to pretrained diffusion (Zhang et al., 2023; Ye et al., 2023); Prompt-to-Prompt, InstructPix2Pix, Animate Anyone, DragonDiffusion, and Add-it support text editing, identity preservation, feature-guided manipulation, or insertion (Hertz et al., 2023; Brooks et al., 2023; Hu, 2024; Mou et al., 2024; Tewel et al., 2025). VECTILE places reference control inside the interaction: a reference specifies how a site responds to its current input, and that response changes what other sites receive. The object of reuse is a local transformation, rather than the observed trajectory or its final image effect.

Geometry and structured generation. GeoVideo, WorldReel, and PE-Field 4D expose geometry, motion, and projected positions to generation (Bai et al., 2025; Fang et al., 2026; Bai et al., 2026); DG-Mesh and Coupled Diffusion address dynamic reconstruction and multi-view consistency (Liu et al., 2025; Alzayer et al., 2026). Correspondence determines where an observation belongs. Our target geometry also determines which input reaches a response after another body changes. It therefore participates in the interaction before its result is projected into a camera, rather than only aligning appearances across views.

Learned response and neural transport. NCLaw already learns constitutive laws from motion, transfers them to new conditions, and couples separately learned materials (Ma et al., 2023). PhysDreamer fits material properties to generated reference motion, while PhysGaussian integrates simulation with Gaussian scenes (Zhang et al., 2024; Xie et al., 2024). GNS and MeshGraphNets learn dynamics; Neural Radiosity, Neural Radiance Caching, and RenderFormer++ learn light transport or organize computation around it (Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021; Hadadan et al., 2021; Müller et al., 2021; Du et al., 2026). VECTILE makes the response an independently replaceable reference instance in generative editing. A shared encoder changes the installed response without refitting its model to each target. The defining operation keeps other responses fixed while recomputing their coupled consequences: replacing the front sheet's reference can change the rear sheet's motion without replacing its reference. The reference thus controls an interaction, not a precomputed trajectory.

Composition and locality. Composable Diffusion combines component distributions; Slot Attention, Neural Interpreters, and object-centric theory study structured representations and routed computation (Liu et al., 2022; Locatello et al., 2020; Rahaman et al., 2021; Wiedemer et al., 2024). ComPhy studies physical counterfactuals, and diffusion analyses relate generalization to conditional dependencies (Chen et al., 2025; Kadkhodaie et al., 2024; Bradley, 2026). Input-conditioned operators also predate this work (Jia et al., 2016). We ask whether independently observed responses can interact in a new target, not only coexist in its output. Locality constrains what a reference replacement changes directly; it does not constrain the reach of its downstream consequences.

3 VECTILE: RESPONSE TILES, TILEFLOW, AND VISUAL SYNTHESIS

3.1 EXTRACT AND INSTALL

Let R_j contain a reference observation and its declared acquisition information. A shared encoder within each domain extracts $r_j = E_\theta(R_j)$. Installation specifies a binding $b(i) = j$ and a target-local frame c_i. The code stays fixed during target rollout, while input, state, and geometry can change. Three edits therefore have distinct meanings: move the source to change the query, reinstall a material to change its frame, or replace the reference to change the response. The other installed tiles need not be replaced when their inputs change.

For active sheets, RGB, observed support, timestamps, and acquisition metadata yield a 32-dimensional response code. For optical reference tasks, observed probes and their visual features yield a 16-dimensional code. Without an independent optical reference, a source-appearance prior supplies an effective response.

3.2 TILEFLOW: FIXED RESPONSES, RENEWED TARGET INPUTS

We roll out the installed response tiles by alternating local evaluation and a target update:

$$(h_i^{k+1}, o_i^k) = P_\theta(r_{b(i)}, u_i^k, h_i^k, c_i^k), \tag{1}$$

$$(W^{k+1}, u^{k+1}) = \mathcal{U}\big(W^k, \{o_i^k\}_i, a^k\big). \tag{2}$$

Here u_i^k is incoming influence, h_i^k optional local state, o_i^k the local output or deformation driver, and a^k the authored source/intervention. The target update uses all local outputs to form subsequent inputs. Stateless optical response omits h; stateful material response retains it. k indexes interaction rounds or physical-time samples, not diffusion steps.

Along a two-site optical path, replacing r_A changes $P_{r_B} T_{B \leftarrow A}(W) P_{r_A} u_A$ without replacing r_B. For active sheets, replacing the front response can change the geometry through which the rear receives

heat. These are consequences of recomputation in the target, not additive combinations of completed image edits. We call this rollout organization TILEFLOW.

3.3 TWO TARGET UPDATES

Optical response and typed transport. Our optical module separates local response from target-dependent propagation. At fixed response code and context,

$$P_r(u, c) = \|u\|_2 f_\theta([u/\|u\|_2, c, r]), \qquad P_r(0, c) = 0, \tag{3}$$

with a defined zero-input branch. For each before/after endpoint, the incoming field is initialized from a projected source s and updated as

$$o^k = P_r(u^k, c), \qquad u^{k+1} = s + T_W(o^k), \qquad u^0 = s. \tag{4}$$

The implementation uses 256 sites, 16 channels, four learned modes, and four local-response evaluations. T_W is geometry-conditioned typed transport; the two endpoints share parameters and installed references. These are signed learned latent fields, not calibrated radiance. Positive homogeneity holds for the local response at fixed conditioning, not for the full image model. It does not establish passivity or an exact multibounce light solver.

Material state, shape, and renewed exposure. A shared response MLP maps the installed code and current temperature/activation to bounded heating, cooling, activation, recovery, bending, and contraction coefficients. Prescribed state updates use these coefficients and incident heat to predict the next local state and deformation drivers. LibShell realizes all bodies under their target constraints (Chen et al., 2018); heat visibility is then recomputed against their joint predicted geometry. This instantiates Eq. 2 through changing shape and exposure, rather than through the optical T_W. It retains history: a warm sheet can respond after current heating disappears. Appendix B.2 gives the state equations and rollout.

3.4 VISUAL SYNTHESIS AT THE BACKBONE'S HIERARCHY

Optical spatial goals and repeated control. Figure 2 shows the implemented optical path. A decoder combines before, after, and difference fields with legal condition features, upsamples a 16×16 site grid to 32×32, and predicts a latent goal g_z, an auxiliary RGB goal g_{RGB}, and spatial features F. The latent goal is a residual over the source latent. Neither goal is the final output image.

A shared spatial controller reads these goals, current output-token features, the noise level, and a detached preceding clean estimate. It adds residuals after Wan blocks 5, 10, 15, 20, 25, and 30. Controller memory runs across these depth locations and resets at each denoising call. Between sampling steps, feedback is the model's own estimate $\widehat{z}_0 = z_t - \sigma_t v_t$ and its frozen-VAE decode. TileFlow and the spatial goal are prepared once per request; generated RGB is not fed back into the optical transport state. Thus interaction depth, spatial upsampling, transformer depth, and denoising time remain distinct.

The synthesis network uses a frozen pretrained Wan backbone with trainable conditioning and low-rank adapters. Repeated spatial control makes the predicted consequence available at successive feature depths, while feedback relates that goal to what the current sample has already expressed.

Material videos and multiple views. The material branch computes one world trajectory $\widehat{W}_{0:T}$. Each requested camera observes that same trajectory and supplies controls to video synthesis:

$$C_c = \mathrm{Observe}_c(\widehat{W}_{0:T}), \qquad Y_c = G_\psi(I_{0,c}, R, C_c). \tag{5}$$

Physics-query conditioning and predicted VACE controls guide video synthesis. A camera change reprojects the trajectory without refitting the response or changing stimulus history. The material interface uses these projected controls, rather than the optical spatial-goal controller.

3.5 LEARNING THE RESPONSE AND ITS VISUAL EXPRESSION

Material response learning uses local transitions, perturbed-state queries, and cross-observation agreement. Video flow loss trains synthesis without differentiating through the shell solve. Optical

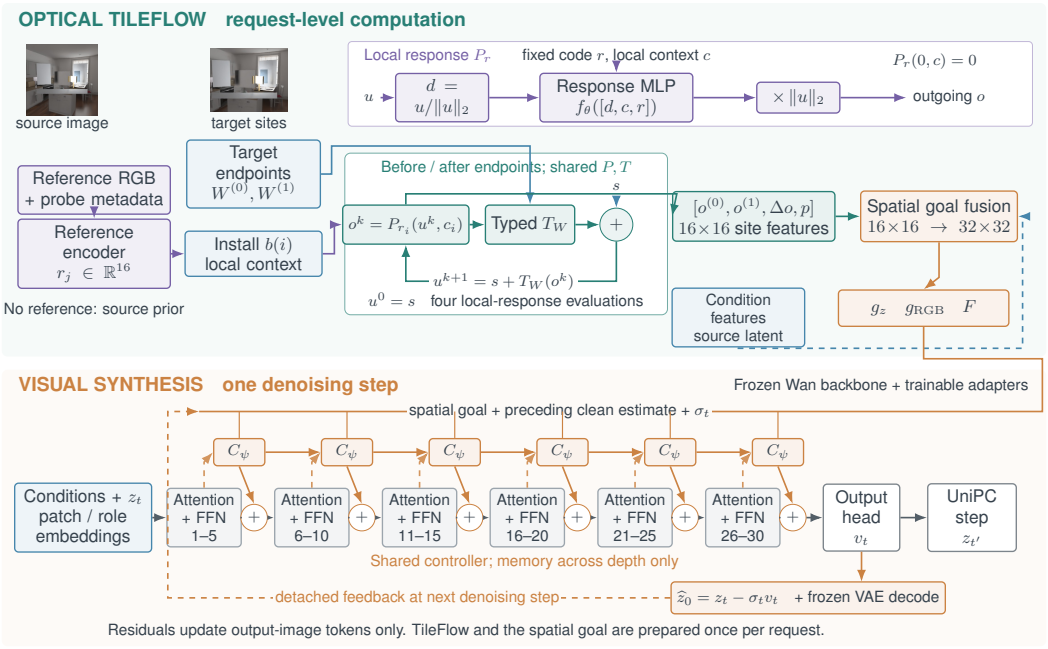

Figure 2: Optical TileFlow and hierarchical visual synthesis. Reference-conditioned local response alternates with target-dependent typed transport. Before/after results form a spatial goal. A shared controller corrects output tokens at six depths of the diffusion transformer, using the preceding denoising estimate as feedback. The pretrained Wan backbone is frozen, with trainable adapters and control modules. No teacher image supplies the goal or feedback at inference.

training warms up the latent/RGB goals, then jointly optimizes flow and goal losses with separate adapter/module learning rates and balanced changed/unchanged regions. Feedback is a detached model estimate. Training uses same-noise-level self-conditioning; sampling uses the preceding step's estimate. Appendix D.4 gives the schedule.

4 Experiments

4.1 Tasks and Baselines

The main material transfer set contains 68 configurations of one held-out base world, each observed from two cameras. It varies state, stimulus, occlusion, installation, and four known response identities. The reference encoder is shared across these identities; the held-out factors are target conditions and combinations, not arbitrary unseen material laws. Separate tests cross two installations with six reference observations and compose independently observed responses, including an ordered pairing excluded from training. The additional joint-edit factorial remains separate from the main held-out count.

The optical development set has 193 requests in 68 groups across seven room categories. It includes directional, branch, relay, and mixed arrangements with movement, color, and intensity edits. The optical evaluation tests receiver activation, emitter placement, and preservation of unaffected content. It includes source-observed response priors and movement requests with fixed independent response references; Fig. 4 distinguishes these inputs. An earlier PASMVS-based multi-view construction supplies 24 material interventions and 288 rendered pairs for studying redistribution, transmission, and wet-surface effects (Appendix C.5).

VIPER-style/Wan follows the reference-alignment, transfer, and adaptation stages using Qwen3-VL-4B, 128 physics queries, and Wan2.1-VACE-1.3B. LightMover-style/Wan retains source, object, movement, color, and exposure conditions. These are local method transfers, not the authors' released checkpoints. Material inference uses 50 Flow Euler steps, shift 3, guidance 1, and a common seed; the

Table 1: Material-task inputs and supervision. Later target state and controls are predictions, not additional future observations.

Quantity	VIPER-style/Wan	VECTILE
Reference	One RGB clip	RGB, support, timestamps, acquisition metadata
Target	Initial image; state/stimulus text	Same image; explicit initial state, geometry, stimulus, binding
Later state	Implicit in generated video	Predicted by TileFlow
Supervision	Reference transfer and video flow	Local transition labels and video flow

Table 2: Multi-view material transfer. Each configuration has two generated views. The 68 configurations share one held-out base world; they are not 68 independent worlds.

Method	Configurations	Correct	Incorrect	Low-visibility
VIPER-style/Wan	68	6	46	16
VECTILE	68	52	0	16

optical comparison uses 30-step UniPC and seed 0. Saved LightMover-style outputs are unchanged. Appendix D distinguishes their checkpoint histories and costs.

4.2 EVALUATION

An edit is correct when the intended local and downstream responses appear while object identity and unaffected content remain usable. A material configuration must succeed in both views. Low-visibility effects remain a separate category: the expected change in the target rendering is too small to distinguish reliably at the displayed scale. This denotes weak observable change, not blurry pixels in the generated output. Intentional rest under cold/no-stimulus controls is correct, not ambiguous. We supplement these judgments with background-compensated generated-RGB optical flow. The seven optical recoveries use the authors' assessment within 69 previously unsuccessful baseline requests, not a new aggregate success rate over all 193 images. Input and supervision differences are explicit (Table 1); this is not a fully matched architecture-only comparison.

5 RESULTS

5.1 SAME RESPONSE, DIFFERENT MOTION

Hold the response fixed; change its target. VECTILE succeeds on 52 configurations, versus six intended-static controls for native transfer (Table 2). A cold, unheated sheet remains still, while a warm sheet can keep responding after heating stops. Reuse must permit a different motion: the response is preserved, but its query carries a different history. Median motion-amplitude ratios are 1.085 versus 0.305, with direction cosines 0.908 versus 0.143, over 124 valid body-view measurements.

Change the observation; preserve the behavior. Two installation axes crossed with six reference-camera/appearance observations give 12/12 correct edits for VECTILE and 0/12 for native transfer. Both have small body-RGB differences across observations: 0.0039–0.0040 and 0.0032–0.0034. Yet the baseline is stable largely because it stays still. This reveals why invariance is insufficient: reuse needs stability across reference observations *and* correct action under a new installation.

5.2 LOCAL REPLACEMENT, COUPLED CONSEQUENCES

Same pair, different consequence. Figure 3 installs separately observed r_{00} and r_{03} on front and rear sheets. The pair is absent from target training and validation, although both components occur elsewhere. Both layouts succeed in both views; across 16 multi-response/joint configurations, 13 are correct in both cameras and three have low-visibility effects. The rear is nearly inactive in one layout but active in the other; changing camera preserves that process. Thus the same responses generate different consequences, rather than replay two reference clips.

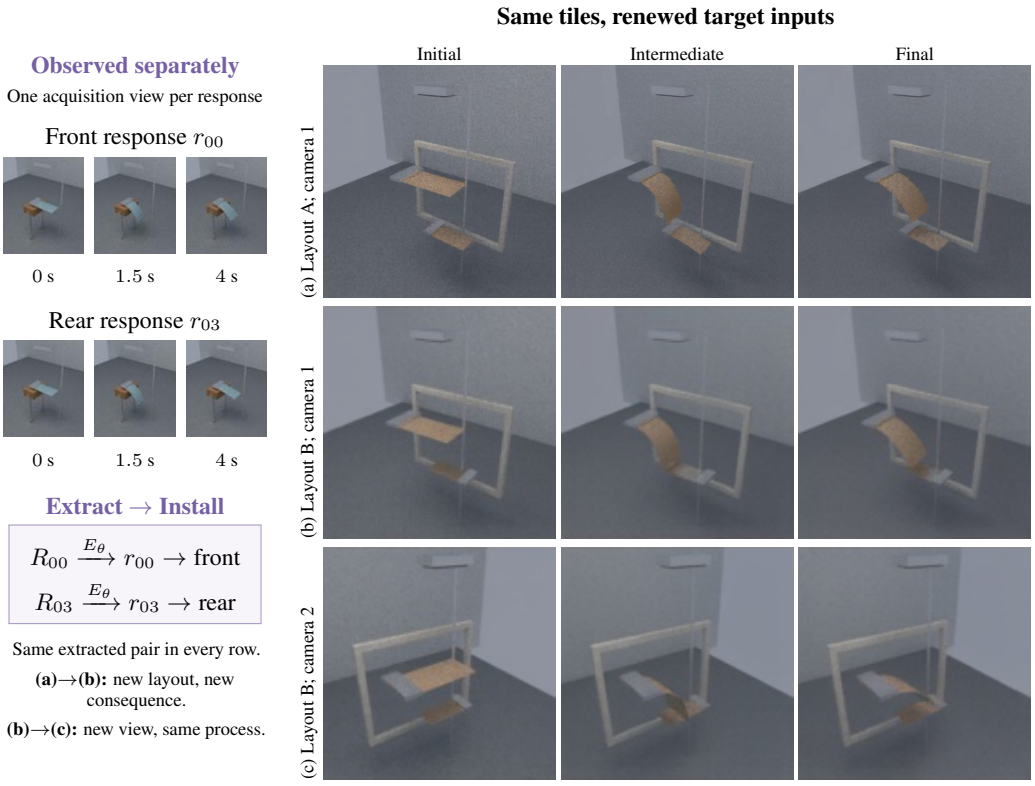

Figure 3: Same response tiles, different consequences. The separately observed front (r_{00}) and rear (r_{03}) responses stay fixed in all rows. (a)→(b): changing the target layout makes the rear responsive. (b)→(c): changing only the camera re-observes the same process. The pair is absent from target training. Left: actual reference snapshots; right: saved generated frames at synchronized times.

Table 3: Reuse and composition success. Fractions are successes / evaluated settings (first row) or generated videos (other rows; one per camera). "—" means unavailable through the native single-reference interface. Native and extended comparisons have different training histories.

Test	Evaluation unit	Comparison success / total	VECTILE success / total
Observation × installation	Setting: 2 installations × 6 observations	Native: 0/12	12/12
Unseen pair, new world	Video: 2 layouts × 2 views	Native: —	4/4
Unseen pair, seen world	Video: 2 layouts × 2 views	Extended: 0/4	4/4
Response/source/install factorial	Video: 8 configurations × 2 views	Extended: 0/16	16/16

Replace one response; recompute the interaction. The factorial pairs front responses r_{00} and r_{02} at two source placements and two rear installations, keeping rear response r_{03} fixed. All 16 views of eight configurations succeed. The replacement strongly suppresses rear motion in one arrangement, changes it less after moving the source, and becomes nearly irrelevant after reinstallation (Table 15). Local replacement can thus have non-local consequences: the unchanged rear response receives a new input. The replacement is reusable; its downstream image difference is not fixed.

The r_{03} factorial is a separate post-primary development analysis. The extended VIPER-style comparison receives multiple slots and target conditions but also fails its unchanged training controls (0/4 versus 4/4); Table 3 and Appendix F retain this comparison context.



Figure 4: A preserved response still needs a new visual consequence. (a) Recoloring uses a source-observed response prior; the missing right-receiver change is restored. (b) Independent response tiles stay fixed while the emitter moves. The ribbon identifies these reference inputs; LightMover-style uses native controls. Targets are evaluation endpoints; outputs are saved generations. Appendix H shows all seven recoveries.

5.3 A REQUESTED CHANGE CAN STILL MISS ITS RECEIVER

The material tests expose the meaning of response reuse; optics tests its visual expression. VECTILE recovers seven baseline failures: one recoloring and six joint moves. In kitchen q17 (Fig. 4a), right-panel mean absolute RGB change is 0.1827 in the target, 0.0087 for LightMover-style, and 0.1526 for VECTILE; the latter two change-direction cosines are -0.129 and 0.975. One baseline receiver follows the blue instruction while the other barely changes. Thus recognizing a global edit does not ensure that every required receiver expresses it.

Recoloring uses a source-image response prior. Movement additionally supplies fixed independent references to VECTILE: the emitter changes position while response identities stay fixed (Fig. 4b). Emitter shape, placement, and visible response improve, with residual spill and fine-geometry errors. These recoveries support a second requirement of reusable behavior: a target-dependent consequence must survive synthesis at its affected sites. Here the combined model expresses fixed-response consequences; the material factorial separately tests response replacement.

5.4 WHAT THE TARGET AND SYNTHESIS MUST CONTRIBUTE

With the response model fixed, zeroing temperature, activation, and conduction reduces world-response successes from 12/12 to 4/12; freezing exposure geometry changes 14/14 to 11/14 (Table 4). The former removes queried and carried history; the latter keeps local updates but prevents their geometry from renewing exposure. This separates two target responsibilities: state carries history, while geometry routes interaction. A tile cannot replace either with a fixed reference motion. These are numerical-world judgments, not additional video scores.

Table 4: Successful evaluations under fixed-model interventions. Fractions give successes / evaluated cases under the full and modified settings. World-response success is judged on the numerical rollout; video and image success is judged on the generated output. Each row uses the same trained model before and after its intervention.

Intervention	Evaluation unit / subset	Full success / total	Modified success / total
Zero temperature, activation, conduction	World-response case; 12 state-sensitive cases	12/12	4/12
Freeze exposure geometry	World-response case; 14 multi-body/joint cases	14/14	11/14
Remove predicted video controls	Generated video; 5 cases	5/5	0/5
Zero field projection (early optical model)	Generated image; 6 cases	0/6	0/6

Removing predicted VACE controls changes five video successes to failures, including background/object artifacts; it also shifts the conditioning distribution. Conversely, disabling the early optical field projection leaves six failures unchanged. These outcomes reveal that containing a transport branch and using its consequences are different achievements: the visual interface must have a task-level effect, not merely change pixels. Appendix G gives the scopes.

5.5 From early failures to target-aligned synthesis

We initially reused a static source-derived optical grid, but moved occluders and reinstalled receivers could leave its relations unchanged. Rebuilding edited endpoints makes the edit change the inputs received. Coarse sampling also omitted visible receivers, motivating explicit site coverage. These failures show why installation must address an actual interaction site: a richer response cannot compensate for a missing site or an unchanged query.

An unedited first image also conflicted with an already edited installation in 24 of 314 configurations. Rendering the edited initial world for both methods resolved this without future states. The edit must enter the state being evolved, not remain only a verbal instruction.

Favorable field diagnostics in PASMVS likewise did not consistently survive synthesis: RGB-only calibration improved consequence error on 12/24 edits but harmed preservation and lost to a stronger jointly fitted comparison. This motivated spatial goals and repeated control. Together, these observations distinguish *addressing the response, renewing its input, and expressing its consequence*. The seven optical recoveries show progress at their interface, without isolating a single cause.

6 Discussion and conclusion

Reference-based control is important for generative models and visual creation. Progress in transferring appearance and behavior brings a deeper question into focus: what semantic unit should a reference carry into a new setting? VECTILE studies reactive behavior—how something responds to its input, state, and surroundings. These responses matter in natural-world editing: preserving a material's behavior should not force it to repeat a motion observed under different conditions.

Our formulation and experiments develop this idea into response tiles and clarify how to extract, install, and reuse them. Material editing demonstrates local replacement with coupled consequences; relighting exposes the complementary need to express those consequences in the image. Together they make a way of responding, rather than a particular outcome, an operational unit of generative control.

This study instantiates tiles within specified interaction models and installation sites. Discovering the reusable units themselves—and the observations needed to identify them—from unconstrained video remains an open direction. We plan to explore richer settings, including characters with distinctive personalities expressed through their reactions to new situations.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices A–B specify the response representations, target updates, and synthesis interfaces. Appendices C–D describe data construction, splits, training objectives, and optimization and sampling settings. Appendices E–H report evaluation procedures, ablations, and per-task results; Appendix I gives the computational environment and complete primary scene outcomes.

AI USE STATEMENT

In this work, we used generative AI tools for generate synthetic data sets, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, propose or refine hypotheses, support qualitative and thematic data analysis, interpret results. And the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for create or modify scientific figures or images, create or edit software code, identify relevant literature. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey, LLM-generated code was verified and tested for correctness by us. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Hadi Alzayer, Yunzhi Zhang, Chen Geng, Jia-Bin Huang, and Jiajun Wu. Coupled diffusion sampling for training-free multi-view image editing. In *CVPR*, pp. 43686–43696, 2026.

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. GeoVideo: Introducing geometric regularization into video generation model. In *NeurIPS*, 38:64252–64272, 2025. doi: 10.52202/085713-1931.

Yunpeng Bai, Haoxiang Li, and Qixing Huang. PE-Field 4D: Video generation models as canvas. In *ACM SIGGRAPH Asia*, 2026.

Arwen Bradley. Local mechanisms of compositional generalization in conditional diffusion. In *ICML*, 2026.

André Broekman and Petrus Johannes Gräbe. PASMVS: A perfectly accurate, synthetic, path-traced dataset featuring specular material properties for multi-view stereopsis training and reconstruction applications. *Data in Brief*, 32:106219, 2020. doi: 10.1016/j.dib.2020.106219. Dataset version 2: 10.17632/fhzfnwsnzf.2.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, pp. 18392–18402, 2023.

Hsiao-Yu Chen, Arnav Sastry, Wim M. van Rees, and Etienne Vouga. Physical simulation of environmentally induced thin shell deformation. *ACM Transactions on Graphics*, 37(4):1–13, 2018. doi: 10.1145/3197517.3201395.

Tianxiao Chen, Hanmo Chen, Huajin Chen, Bo Li, Qi Ye, and Peng-Tao Jiang. VIPER: Visual in-context physics reasoning for physically plausible video generation. *arXiv:2607.23472*, 2026.

Zhenfang Chen, Shilong Dong, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B. Tenenbaum, and Chuang Gan. Compositional physical reasoning of objects and events from videos. *IEEE TPAMI*, 47(9):7689–7703, 2025. doi: 10.1109/TPAMI.2025.3574322.

Huangsheng Du, Haoran Zhu, Youcheng Cai, Jinyang Meng, and Ligang Liu. RenderFormer++: Scalable and physically grounded feed-forward neural rendering. *arXiv:2606.30380*, 2026.

Shaoheng Fang, Hanwen Jiang, Yunpeng Bai, Niloy J. Mitra, and Qixing Huang. WorldReel: 4D video generation with consistent geometry and motion modeling. In *CVPR*, pp. 11195–11206, 2026.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *13th Scandinavian Conference on Image Analysis*, LNCS 2749, pp. 363–370, 2003.

Saeed Hadadan, Shuhong Chen, and Matthias Zwicker. Neural radiosity. *ACM Transactions on Graphics*, 40(6):236:1–236:11, 2021. doi: 10.1145/3478513.3480569.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, pp. 8153–8163, 2024.

Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *NeurIPS*, 29:667–675, 2016.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *ICLR*, 2024.

Isabella Liu, Hao Su, and Xiaolong Wang. Dynamic gaussians mesh: Consistent mesh reconstruction from dynamic scenes. In *ICLR*, 2025.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Composable visual generation with composable diffusion models. In *ECCV*, 2022.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 33, 2020.

Pingchuan Ma, Peter Yichen Chen, Bolei Deng, Joshua B. Tenenbaum, Tao Du, Chuang Gan, and Wojciech Matusik. Learning neural constitutive laws from motion observations for generalizable PDE dynamics. In *ICML*, PMLR 202, pp. 23279–23300, 2023.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. DragonDiffusion: Enabling drag-style manipulation on diffusion models. In *ICLR*, 2024.

Thomas Müller, Fabrice Rousselle, Jan Novák, and Alexander Keller. Real-time neural radiance caching for path tracing. *ACM Transactions on Graphics*, 40(4):36:1–36:16, 2021. doi: 10.1145/3450626.3459812.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *ICLR*, 2021.

Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation*. MIT Press, fourth edition, 2023.

Nasim Rahaman, Muhammad Waleed Gondal, Shruti Joshi, Peter V. Gehler, Yoshua Bengio, Francesco Locatello, and Bernhard Schölkopf. Dynamic inference with neural interpreters. In *NeurIPS*, 34, 2021.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *ICML*, PMLR 119, pp. 8459–8468, 2020.

Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models. In *ICLR*, 2025.

Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. In *ICLR*, 2024.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-integrated 3D gaussians for generative dynamics. In *CVPR*, pp. 4389–4398, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv:2308.06721*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pp. 3836–3847, 2023.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3D objects via video generation. In *ECCV*, pp. 388–406, 2024. doi: 10.1007/978-3-031-72627-9_22.

Gengze Zhou, Tianyu Wang, Soo Ye Kim, Zhixin Shu, Xin Yu, Yannick Hold-Geoffroy, Sumit Chaturvedi, Qi Wu, Zhe Lin, and Scott Cohen. LightMover: Generative light movement with color and intensity controls. In *CVPR*, pp. 8997–9007, 2026.

A RESPONSE TILES AND THE SCOPE OF TILEFLOW

A.1 VOCABULARY AND LEVELS OF ABSTRACTION

A *transport reference* is the reusable local response specified by reference observations. A *response tile* is its extracted representation within a trained model. The tile is installed using a separate binding and target-local frame. It need not correspond to a rectangular image crop, and moving or reusing it does not mean transporting a cached output image.

Extract obtains the response representation; **Install** associates it with target sites; **TileFlow** repeatedly evaluates the installed responses as the target renews their inputs; **Visual Synthesis** expresses the resulting consequence in generated images or videos. We use "roll out the installed response tiles" for the computation performed by TileFlow. The name is local terminology for that computation, not a synonym for diffusion flow matching.

These stages distinguish a representation, a mechanism, and its demonstrations. The representation separates response identity from the target consequence. The mechanism is local response followed by target-mediated renewal. Optical and active-material tasks demonstrate that organization using different physical quantities, neural modules, and synthesis interfaces. The shared encoder is shared across references *within a domain*; the 16-dimensional optical and 32-dimensional material codes are not a common cross-physics latent space.

Table 5: One organization, two implementations. The common architecture does not identify the optical and material state variables or synthesis modules.

Stage	Primary optical implementation	Stateful material implementation
Extract	Reference-probe features $\rightarrow$ 16-D law; source prior when no independent reference	Timed RGB/support/metadata encoder $\rightarrow$ 32-D response
Install	Per-site reference index; target local context and learned mode organization	Per-body/face response; target mounting, material axis and constraints
TileFlow	$o = P_r(u, c)$; $u \leftarrow s + T_W(o)$	Local state/driver update; constrained shape solve; recomputed heat exposure
Visual Synthesis	Spatial goal; shared depth-wise controller; preceding-denoising-estimate feedback	Shared world trajectory; camera-specific renders; predicted VACE controls

A.2 OPERATIONAL IDENTITY AND REUSE

The code coordinates are not physical measurements. For a specified set of local queries $\mathcal{Q}$, write

$$r \sim_{\mathcal{Q}} \widetilde{r} \quad \Longleftrightarrow \quad P_\theta(r; q) = P_\theta(\widetilde{r}; q) \quad \text{for every } q \in \mathcal{Q}. \tag{6}$$

For approximate models, compare outputs on those queries rather than requiring identical latent coordinates. An invertible code transformation $\widetilde{r} = g(r)$ with $\widetilde{P}(\widetilde{r}, q) = P(g^{-1}(\widetilde{r}), q)$ leaves every rollout unchanged. Our experiments reuse such operational responses within the stated active-material family; they do not identify an arbitrary material law from one ambiguous observation.

The architecture prescribes the response/target split and learns within it. A generic function of (r, u, h, c) with the same independent binding and target-update loop can implement this factorization. Optical homogeneity, a particular typed carrier, grid resolution, and a spatial controller are choices of implementation, not the definition of a reusable response. H2 is selected as the primary optical parameterization because it makes the local-response/propagation boundary explicit. H1 remains an alternative parameterization, not a second canonical definition of VECTILE.

A.3 RELATION TO RADIATIVE TRANSPORT

For fixed linear optical operators, $U = S + TPU$ gives $U = S + TPS + (TP)^2 S + \cdots$ when that series converges. This motivates alternating local transformation and propagation through the target (Pharr et al., 2023). Our implemented optical response is nonlinear and uses a finite rollout.

No infinite-series convergence, passive light solver, or radiometric calibration is claimed. For the active material, recurrence is through local state, constrained shape, and subsequent heat exposure, not repeated optical scattering.

A change of output camera occurs after material rollout. A change of source or installation occurs within target construction and alters the received input or local frame. Replacing a response changes the transformation at its bound sites. These distinct operations are not interchangeable forms of reference-image conditioning.

B Response models and visual-synthesis interfaces

B.1 Material reference extraction

The scene response model uses a timestamp-aware visual encoder. Each reference view is resampled to 13 actual observation timestamps and resized to 48×48. A 5×5 stride-2 convolution maps RGB to 16 channels; a 3×3 stride-2 convolution produces 32 channels. Adaptive 4×4 pooling and a $512 \to 48$ projection give a frame feature. Convolutions and projection use SiLU. Appending time divided by four gives a 49-dimensional sequence input to a bidirectional GRU with 48 hidden units per direction. A $96 \to 32$ projection and temporal mean produce a visual code.

The support branch measures masked centroid, area, second moments, and mean brightness relative to the first frame. The 78 values across 13 frames pass through a $78 \to 96 \to 32$ MLP. A $10 \to 32 \to 32$ acquisition branch has a zero-initialized final layer. It receives normalized camera-to-look-at direction, field of view divided by 90, reference source position and strength, and source on/off times divided by four. The branches are added and averaged across supplied reference observations. These are observed reference quantities, not target-future geometry.

The small response encoder and the Qwen branch serve different roles. The small encoder extracts the response used by the material rollout. Qwen supplies inherited generative conditioning. Both remain in the reported scene implementation; they should not be drawn as a single shared feature extractor.

B.2 Material local response and state update

For each face, the 32-dimensional response code is concatenated with current temperature θ and activation m. A SiLU MLP of widths $34 \to 96 \to 96 \to 6$ predicts six sigmoid-bounded coefficients (Table 6). They depend on current local state rather than forming one fixed coefficient vector regressed once per reference.

Table 6: Learned active-material coefficient bounds. These are engineering priors for the constructed process; temperature and activation are dimensionless.

Coefficient	Role	Permitted interval
α	Incoming heat response	$[0.4, 1.5]$
β	Cooling rate	$[0.3, 1.2]$
τ^+	Activation time scale	$[0.12, 0.77]$
τ^-	Recovery time scale	$[0.3, 1.3]$
κ	Curvature scale	$[1.0, 3.2]$
γ	Contraction scale	$[0.005, 0.055]$

With incident heat q_i, time step Δt, and face-adjacency Laplacian L, the implemented state equations are

$$\theta_i^+ = \max\{0, \theta_i + \Delta t[\alpha_i q_i - \beta_i \theta_i + 0.015(L\theta)_i]\}, \tag{7}$$

$$e_i = 1 - \exp[-\max(\theta_i, 0)], \qquad \tau_i = \begin{cases} \tau_i^+, & e_i > m_i, \\ \tau_i^-, & e_i \le m_i, \end{cases} \tag{8}$$

$$m_i^+ = e_i + (m_i - e_i)\exp(-\Delta t/\tau_i), \tag{9}$$

$$b_i^+ = \kappa_i m_i^+, \qquad c_i^+ = \gamma_i m_i^+. \tag{10}$$

The equilibrium e_i uses current θ_i, not θ_i^+. Here b_i^+ and c_i^+ denote curvature and contraction drivers; c_i^+ in these material equations is not the target-frame symbol c_i^k in Eq. 1. The update equations, adjacency, and mechanical solver are specified; the reference-conditioned model predicts the coefficients.

Mechanical update and exposure. The target provides rest vertices, faces, fixed vertices, material axis, thickness, initial temperature and activation, stimulus, and slot bindings. Curvature and contraction alter the shell's rest forms following Chen et al. (2018); LibShell solves the constrained geometry. Every body updates using its own response before incident heat is recomputed from all updated bodies. A rear body neither copies a reference trajectory nor reads a ground-truth front-body future.

The trajectory has 49 samples over four seconds. Seventeen frames are rendered at 256×256, 16 samples per pixel, for the video-control cache. The first control frame supplies the common legal initial-image condition; VACE reactive channels receive the predicted remainder. The final videos are full-noise diffusion outputs, not these renders. Video training does not differentiate through the native LibShell solve.

B.3 Optical extraction, installation, and typed response

H2 replaces the learned-graph interaction module with conditional local response and typed transport, while retaining the common goal and synthesis interface. Default width is 64, with 16 channels and four modes in the typed field. The existing implementation also accepts 12 channels with a feature projection; that option is not the default H2 configuration. Four response evaluations are performed per endpoint.

Independent references. The current input format supports four references, each with four known probes. Four visual contact-sheet latent frames occupy the final condition positions. Each sheet is split into four probe patches. A two-layer 16-to-64-channel convolutional encoder and 4×4 pooling produce 16 tokens per probe. Numeric observed/control features, within-probe coordinates, and probe identity are encoded as metadata and added to image features. Learned key/value projections give 64 tokens per reference. These counts are properties of the present input format, not a claim of arbitrary-cardinality support.

H2 pools reference tokens with learned scalar attention, respecting supplied probe support. A $64 \to 64 \to 16$ MLP with a final tanh gives the reference law. Binding gathers one code for each target site. A requested binding to an entirely unsupported reference raises an error. The same extracted reference code is available to both endpoint rollouts; target queries do not change the extracted independent code.

Source-prior path. For unbound sites, a two-layer convolutional encoder reads only the source-image latent, pools to 16×16, and combines those features with the supplied colors in a $67 \to 64 \to 16$ prior-law MLP. The general five-condition encoder is separate. Its outputs help the spatial-goal decoder, but do not replace independent law extraction. Without a world packet, no transport is called: the structure falls back to source appearance. Such outputs are not evidence of target-world transport execution.

Local context and recurrence. Each endpoint contains local features of width 7, a width-7 source, pairwise edge features of width 9, and validity. Absolute position entries are zeroed in the local-context MLP; normal and local size are retained. Geometry remains available to target transport. A bias-free linear map projects the source to the field width. The local response is Eq. 3, using a $(d + d + 16) \to 128 \to d$ SiLU MLP for $d = 64$ by default.

For nonzero u, normalization uses its actual norm, rather than an epsilon floor; the zero-input branch divides by one before multiplying by zero. This gives exact zero response and positive homogeneity at fixed context and code, apart from floating-point effects. It does not imply additivity, nonnegative energy, conservation, or a homogeneous full generator.

The geometry-conditioned typed transport operator uses receiver–sender edge indexing, with a receiver-to-sender vector. The inherited 256-site packet uses valid-count normalization, not the older

small-world normalization. The local response returns a signed field reshaped to channels $\times$ modes; target transport acts on that field, and the source is reintroduced for the next call.

The loop records four outgoing states and returns the fourth. It also computes a next-incident state after each outgoing state, including the final one. That final next-incident state is not consumed by another local response in the returned image features. Consequently, "four rounds" names four local-response evaluations, not four physically certified bounces or four propagated links before the returned state.

Two-world re-execution. Before and after endpoints are processed separately with shared parameters. The decoder receives $[o^{(0)}, o^{(1)}, o^{(1)} - o^{(0)}, p]$, where p is a condition prior. This is two-world re-execution and feature differencing, not a signed perturbation-source algorithm. The authored endpoint geometry is fixed during the optical diffusion trajectory.

B.4 Spatial goals and depth-wise control

Five native VAE conditions, each with 16 latent channels, are concatenated as 80 channels and encoded by 3×3 convolutions $80 \to 96 \to 64$. Adaptive pooling gives 256 site features. The concatenated before/after/difference/prior feature has 256 channels on a 16×16 grid. A convolutional decoder and factor-2 PixelShuffle produce 64-channel 32×32 features, fused with resized legal-condition features.

A 16-channel convolution predicts a residual over the source latent resized to 32×32; the resulting latent goal is resized to the actual latent grid. Its final layer is zero-initialized. A separate 3-channel sigmoid head produces the 32×32 RGB goal. This RGB head is not the VAE decode of the latent goal, and neither head is the final generated image. The intermediate hierarchy is explicitly spatial, not a learned Fourier-frequency decomposition.

Controller input. At six target-stream locations, the controller receives a 1,536-channel host feature projected to 64 channels; 64 spatial goal features; the 16-channel latent and 3-channel RGB goals; preceding latent/RGB estimates; the corresponding goal-minus-estimate differences; normalized timestep $t/1000$; and a feedback-availability mask. The concatenation has 187 channels. Missing feedback is masked, rather than replaced by the teacher target.

A convolutional gated recurrent unit computes a 64-channel control state. A zero-initialized, bias-free 1×1 convolution maps that state to a 1,536-channel residual. The residual is added only to output-image tokens after blocks 5, 10, 15, 20, 25, and 30. The controller parameters are shared across these sites. Its hidden state runs across transformer depth *within* one forward call and resets for the next diffusion step; only the detached clean estimate and decoded RGB persist between sampling calls.

Inherited path. Both H1 and H2 retain the inherited optical conditioning pathways. The original response core remains frozen while computing before/after/difference features; its field projection remains trainable during joint optimization. Type/role embeddings, native condition projections, extra world/reference token projections, and inherited trainable adapters are also preserved. The legacy field enters output tokens before the first Wan block. Figure 2 focuses on the response, transport, and synthesis-control stages; these pre-existing conditioning details remain part of its input embedding.

A no-legacy hierarchy would be a different configuration. No such result is substituted for the reported H2 checkpoint. Legal condition skips, inherited field injection, and the synthesis backbone offer paths in addition to the new typed rollout, so image success alone does not isolate the causal contribution of that rollout.

Three recurrent processes. Optical TileFlow computes interaction states at request preparation. The spatial controller recurs across transformer depth. Denoising feedback connects completed sampling estimates. These are three distinct recurrences. In the material branch, TileFlow instead evolves physical-time state and geometry before camera readout. The optical feedback arrow should not be added to the material architecture diagram.

B.5 Inference and cache behavior

The sampler constructs legal condition latents and the world packet, prepares TileFlow and its spatial goal once, and starts output latents from independent Gaussian noise. At each UniPC step it predicts velocity, forms $\widehat{z}_0 = z_t - \sigma_t v_t$, decodes that estimate using the actual frozen VAE, and supplies the detached estimate to the next step. All 30 configured steps perform a feedback decode. Teacher values may coexist in a loaded cache packet, but are not passed as model inputs; target images are constructed for evaluation only after generation.

Prepared requests are reused in evaluation only when tensor identity/version and model parameter versions match. Model updates invalidate the cache. The sampler stores checkpoint, manifest, source-code, seed, step-count, and member-list fingerprints; resumption requires the same fingerprint. It does not share controller or prepared-request state across examples.

C Data construction and generalization splits

C.1 Reference library and observation conditions

Four teacher responses define the constructed active-material family (Table 7). Each is observed in two reference rigs. The expanded observation protocol crosses camera azimuths $-40°, 0°, 40°$ with two appearance settings, giving six observations of the same process. Physical timing is unchanged; appearance is not assigned one-to-one to camera identity. These coefficients generate supervision and are not passed to the reference encoder at inference.

Table 7: Teacher responses for the active-material family.

Response	α	β	τ^+	τ^-	κ	γ
r_{00}	1.20	0.55	0.18	0.95	2.40	0.030
r_{01}	0.70	0.90	0.60	0.38	1.35	0.018
r_{02}	0.65	0.45	0.55	1.10	1.85	0.040
r_{03}	1.00	1.10	0.22	0.45	2.80	0.024

C.2 Target state and installation

Target construction separates response identity from state and stimulus. Table 8 lists the six basic conditions before additional occlusion cases. Installation varies fixed side, material axis ($0°/90°$), and world yaw ($-30°/+30°$). Geometry is changed before rollout, not by rotating a completed motion sequence. Occluded cold and warm targets separate absent current heating from retained local state.

Table 8: State/stimulus conditions. The usual on-time is 0.25 s. A zero-strength source provides no heat regardless of its stored off-time.

Condition	Initial θ	Initial m	Strength	Off-time (s)
Cold, no stimulus	0	0	0	1.60
Cold, weak stimulus	0	0	0.35	1.60
Cold, full stimulus	0	0	1.00	1.60
Warm, stimulus removed	0.40	0.35	0	1.60
Warm, short pulse	0.40	0.20	1.00	0.75
Cold, long pulse	0	0	1.00	2.80

C.3 Actual training and evaluation membership

The primary scene evaluation uses one unseen base world, `world_933000`, with 68 configurations and two output views. All four response identities occur in training. Held-out status refers to this

target world and its condition instances, not to new constitutive families. Multi-response and joint-edit studies remain separate.

The final native video-training manifest contains 272 view records, the VECTILE scene manifest 328, and an earlier extended-generator manifest 312. The local response model uses 184 body/slot records. These are distinct implementation histories; they are not interchangeable names for one shared training split. Recipe pools are not counted as completed training data.

Table 9: Factors varied in scene evaluations. Counts distinguish physical configurations from generated observations.

Setting	Size	Held-out or changed factor
Single-response transfer	68 configurations; 136 videos	Unseen base world; known responses and varied state, stimulus, installation
Installation/observation	2 installations $\times$ 6 observations	Reference camera/appearance crossed with target installation
Multi-response/joint	16 configurations; 32 videos	Known and unseen ordered pairings; joint edits in the held-out world
Seen-world new pair	2 layouts; 4 videos	Ordered pair $00 \rightarrow 03$ in a previously used world
Additional factorial	8 configurations; 16 videos	Response, source and receiver installation crossed after primary evaluation

C.4 UNSEEN ORDERED PAIRING AND ADDITIONAL FACTORIAL

Reference observations in Fig. 3. The front and rear use response instances r_{00} and r_{03}. The displayed snapshots come from their actual reference observations at 0, 1.5, and 4 seconds, using the 0° acquisition camera and the first appearance setting. That acquisition is one of the six observations associated with each response in both target layouts. Identical crops isolate the sheet and stimulus in the reference previews; the right-hand panels retain the generated target frames. Reference observations and target sequences therefore have different roles: one supplies the response instance, while the other shows its target-dependent outcome.

The four pairings are $(00, 01)$, $(02, 03)$, $(00, 03)$, $(02, 01)$, with the first response assigned to the front body. Pair $(00, 03)$ is excluded from target training and validation. Both components occur in their respective roles in other pairings. The recorded check finds neither that pair nor its reverse in five relevant histories, including final native and VECTILE manifests. The test is new composition of known responses, not unseen-law acquisition. Some factorial incident/state and direction ranges extend beyond marginal training ranges, so it is not pure interpolation in every local variable.

The original joint set uses rear response r_{01}. The later factorial uses r_{03} to exercise the held pairing, in seen world `world_931000`. It is a post-hoc study, not added to the primary held-out total. Section F preserves this distinction.

C.5 OPTICAL DATA AND TEACHER INTERPRETATION

The earlier PASMVS-based multi-view study uses armadillo, bunny, dragon, and teapot assets (Broekman and Gräbe, 2020). Four object configurations, three intervention families (material redistribution, transmission/extinction, and dry-to-wet surface changes), two variants, and twelve cameras yield 24 world interventions and 288 source/target pairs. Figure 5 and the PASMVS thumbnails in Fig. 1 show actual rendered endpoints. Corresponding source/target cameras are identical, and independent repeat renders estimate difference noise. The examples illustrate scene and intervention diversity, not H2 prediction quality.

The historical development fit uses three source-anchor cameras and a fourth development camera for validation. A matched RGB-only calibration has consequence MSE 0.00266207 versus 0.00234318 for generic fusion, winning on 12/24 edits; a stronger jointly fitted generic comparison has MSE 0.00195487. Preservation error is 0.00016325 versus 0.00013933 for the matched comparison. Thus partial consequence improvements do not establish overall superiority. These development diagnostics remain distinct from the relighting benchmark and are not included in its H2 outcomes.

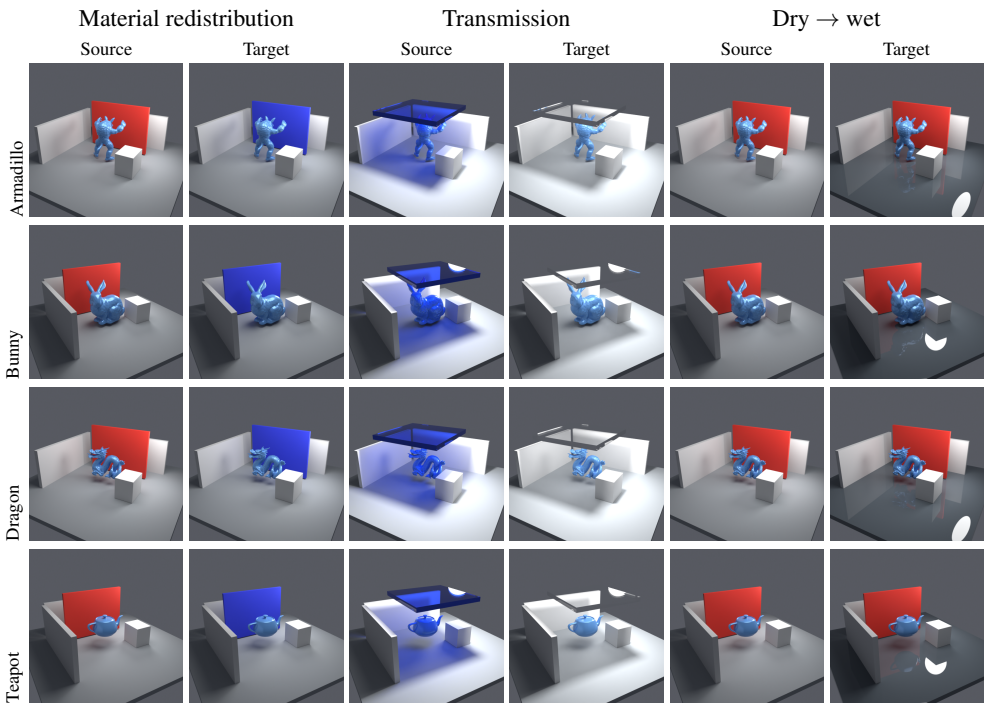

Figure 5: PASMVS-based transport counterfactuals. Actual source/target renders from all four object configurations, shown under three intervention families. Each pair shares its camera and geometry; material changes alter reflections, transmitted light, or neighboring-surface appearance. These are dataset endpoints from the earlier multi-view study, not H2 predictions or additional H2 successes. The full construction comprises 24 interventions observed by 12 cameras.

Response inputs in Fig. 4. The kitchen recoloring request uses authored target geometry with the source-appearance response prior. The living-room emitter-movement request supplies independent response observations as well as authored geometry. Its references remain fixed: the intervention moves the emitter, not a response binding. The native LightMover-style comparator retains its original controls and does not receive these extra references. The diagrammatic input ribbon states this pathway; it is not a panel of measured probe images. Reference replacement is a separate operation, evaluated in the reference/installation set below. These protocols distinguish reference-conditioned reuse from the source-observed specialization of the same response/target separation.

The optical study contains 193 images per method grouped into 68 editing groups across seven room categories. Directional, branch, relay, and mixed arrangements vary source movement, color, and intensity. The set was inspected during development. A separate 100-image reference/installation set tests operations beyond the native five-condition interface; six native-compatible images overlap with the 193 and are not additional independent samples.

A fixed-response cross-world movement construction has 21 cases: 15 training and six held-out. Held-out positive changes are mainly direct-light gains; two negative changes include indirect-light removal. This tests changing global consequences under a fixed local rule, not an isolated forward-multibounce effect.

The optical cache/display path combines separate ambient and controlled-light components. For a fixed endpoint, target construction uses ambient plus the light component multiplied by the requested RGB gain and 2^{EV}, with a common source-derived exposure/display transform. Geometry-changing requests select the authored target endpoint. This formula does not by itself prove the physical accuracy of every underlying light pass, and the available evaluation does not establish a universal no-spill rule. Receiver masks locate evaluation regions; legitimate weak indirect changes must be judged against the actual target and task, not declared erroneous solely because they occur outside a named receiver.

The original four-case real LightMover protocol supplies only source RGB, source mask, and target box. It has no authored world and does not exercise the original structured-world path. It is distinct from both the synthetic hierarchy study and the earlier GT-supervised Stage-3B replay implementation.

D Optimization, named baselines, and cost

D.1 Training the material local response

For target $y = (\theta^+, m^+, b^+, c^+)$, the per-query loss is

$$\ell(\widehat{y}, y) = 20\|\widehat{\theta}^+ - \theta^+\|^2 + 40\|\widehat{m}^+ - m^+\|^2 + 2\|\widehat{b}^+ - b^+\|^2 + 200\|\widehat{c}^+ - c^+\|^2. \tag{11}$$

Each term averages over sampled time/face queries. The adopted response model is trained from a clean initialization for 8,000 updates on 184 body/slot records. AdamW uses learning rate 3×10^{-4}, gradient clipping at 1, and seed 230923. Validation occurs every 100 updates, with the lowest recorded validation loss selecting the response checkpoint.

Training perturbs temperature with Gaussian standard deviation 0.08, activation by 0.04, and incident heat multiplicatively by 0.15, then clips to valid ranges. The specified teacher update provides labels; the extra conduction input is zero for these perturbed queries. One reference camera/appearance observation is also sampled and evaluated on the same queries as the combined-reference code. Its additional loss has weight 0.25, combining single-observation supervision and agreement with the detached combined prediction at relative weight 0.5. This constrains response outputs, not code coordinates.

D.2 Native VIPER-style/Wan and scene synthesis

The native implementation uses frozen Qwen3-VL-4B-Instruct. A set of 128 trainable physics queries traverses 36 language layers; a three-linear-layer connector ($d \to 1024 \to 1024 \to 4096$, GELU) projects outputs into Wan context. Frozen visual prefixes may be cached, but trainable query states are recomputed each update. Query/connector gradients are recorded through the generator.

Native inputs are one RGB reference clip, a legal edited initial image, and target text specifying cold/warm state and source on/off times. Numerical state, reference masks, camera vectors, per-slot binding, local-response labels, and predicted trajectories are absent from this native interface. During transfer/adaptation, one of six RGB observation variants is sampled. Completed stages have 600 alignment, 600 transfer, and 1,000 generator-adaptation updates. Stage 1 uses spatial/photometric self-reference augmentation; stage 2 uses independent references; stage 3 updates queries/connector with rank-8 unscaled self/cross-attention LoRA, including output projections but no feed-forward LoRA.

The Vectile scene generator retains the query/connector route and adds independently identified response slots, target conditions, and predicted VACE controls. Its history includes 200 early alignment, 600 expanded alignment, 600 transfer, 1,000 adaptation, and 100 final blocked-stimulus continuation updates. Local response training is the separate 8,000-update stage above. The earlier target-conditioned VIPER extension has its own history and selected full-v3 checkpoint; it is not pooled with the native baseline.

The scene host is Wan2.1-VACE-1.3B with its conditioning branch active. Final outputs have 17 frames, 256×256, displayed at 4 frames/s. Sampling starts from Gaussian output noise with 50 Flow Euler steps, shift 3, guidance 1, and seed 0. The initial image is rendered after the authored source/installation edit using initial geometry, not a future state.

The published VIPER protocol's higher guidance was tested in the local transfer. Euler/UniPC combinations at guidance 6 caused severe artifacts, including variants preserving physics tokens in the negative branch. Guidance 1 was selected for both scene methods before final evaluation. This is a documented local adaptation, not equivalent capacity or benchmark reproduction of the original larger system.

Table 10: Scene method interfaces. "Computed" quantities are predictions, not target-future inputs.

Quantity	Native VIPER	Extended VIPER	VECTILE
Single RGB reference	Yes	Yes	Yes
Multiple reference slots	No	Yes	Yes
Reference support/timestamps	No	Yes	Yes
Reference acquisition metadata	No	Yes	Yes
Legal edited initial image	Yes	Yes	Yes
Ordinary condition text	Yes	Yes	Yes
Numeric authored conditions	No	Encoded	Text + rollout
Local response supervision	No	No	Yes
Evolving state/geometry	Implicit	Implicit	Computed
Predicted VACE trajectory	No	No	Computed
Target-future trajectory input	No	No	No

D.3 LIGHTMOVER-STYLE/WAN AND THE ORIGINAL OPTICAL MODEL

Five independently VAE-encoded visual conditions—source RGB, object crop, movement map, color, exposure—are concatenated with the noisy output as patch-token groups. Type/role embeddings and native position encoding distinguish conditions and output. Training uses a flow objective; sampling starts from Gaussian noise, without final image pasting or source-initialized suffix refinement. The VACE side branch is inactive for this optical pathway.

The recorded native optical continuation reaches step 1,800. The adopted original VECTILE optical checkpoint is also numbered 1,800 after earlier response/world-condition training. Equal labels are not equal total training. Their final continuation manifests contain 341 and 421 records, respectively. Later unsuccessful continuations remain in cost history. Final optical outputs use 30 steps and seed 0. The earlier shared task-trained optical model at step 4,000 is an ancestor, not the later native-primary checkpoint merely because both are LightMover-style/Wan.

The original VECTILE field path used 12 channels and four local sectors, four interactions, before/after/difference concatenation, elementwise asinh, and a zero-initialized $144 \to 1536$ projection into output-patch embeddings. That describes the inherited field branch, not the new H2 structure with its default 16 channels and spatial-goal/controller path.

D.4 HIERARCHICAL OPTICAL TRAINING AND STEP ACCOUNTING

The hierarchical optical recipe initializes from the designated parent optical checkpoint at step 1,800 and discards the parent's optimizer. It specifies 1,000 structure-only warmup updates followed by 1,200 joint updates. These are new optimizer updates, not the absolute checkpoint label. The accounting is

$$\texttt{update} = n_{\text{warmup}} + n_{\text{joint}}, \qquad \texttt{step} = 1800 + \texttt{update}. \tag{12}$$

Thus the scheduled preview pause after warmup plus 200 joint updates has label 3,000, and the planned full endpoint has label 4,000. The phase counter counts updates within the current phase. The trainer records these counters in each checkpoint; the sampler identifies its input checkpoint automatically.

Warmup trains only the new response-and-goal branch with AdamW, learning rate 10^{-4}, weight decay 0.01, and accumulation 1. Joint training creates a new grouped AdamW optimizer: inherited adapters use 10^{-5}, new hierarchy modules 10^{-4}, and accumulation 2. Gradient clipping is 1. The default seed is 250925, and checkpoints are saved every 200 new updates. Resume restores the optimizer, random-number states, and counters.

Data mixing. Warmup cycles through structured train, structured train, and reference pools. Joint microbatch slots cycle through real, general, structured train, structured train, general, and reference pools. Only rows explicitly marked training are sampled. World-bearing examples use both flow and goal losses; no-world general/real examples receive flow loss only. The structure path is not supplied a fabricated world target for those examples.

Goal and regional objectives. The target latent and its actual frozen-VAE decoded RGB provide training supervision. Changed and unchanged masks are computed from the decoded target–source difference at thresholds 0.04 and 0.01, respectively, then area-resampled. Let $\mathcal{B}(e)$ combine 0.25 times the full error mean with 0.75 times the mean of nonempty changed/unchanged region means. The goal loss is

$$\mathcal{L}_{\text{goal}} = \mathcal{B}\big((g_z - z_*)^2\big) + 0.25\mathcal{B}\big(|g_{\text{RGB}} - D(z_*)_{32}|\big). \tag{13}$$

The RGB target is resized to 32×32 after actual VAE decoding; the latent target is not silently resized to force agreement. World-bearing joint examples add $0.25\mathcal{L}_{\text{goal}}$ to a flow objective mixing original cache-weighted flow and shared equal-entity regional flow at equal weight. Regional masks are explicitly training-only maps. No evaluation mask enters generation.

Self-conditioning versus sampling feedback. For native shifted sigma σ, training uses $z = (1-\sigma)z_* + \sigma\epsilon$ and target velocity $\epsilon - z_*$. With probability 0.5 on world-bearing microbatches, a no-gradient first prediction at the same sigma supplies $\widehat{z}_0 = z - \sigma v$ and a frozen-VAE RGB decode. The gradient-bearing prediction sees this detached feedback. Sampling instead sees the preceding completed step's clean estimate. This difference is retained in the method description; the training is not an on-policy multi-step unroll or teacher-residual feedback.

D.5 Recorded computational cost

Table 11: Original recorded training and inference costs. These totals predate H1/H2. Traced training includes listed ancestors and discarded continuations; optical totals exclude the earlier shared host whose elapsed time is unavailable. Sampling excludes loading and most preprocessing.

Domain	Method	Training (s)	Outputs	Median sample (s)
Scene	Native VIPER-style	9,066.849	136 videos	5.931
Scene	VECTILE	15,708.374	136 videos	6.496
Light	Native LightMover-style	1,299.101	193 images	2.913
Light	Original VECTILE	6,294.134	193 images	7.416

H2 adds request preparation, six controller calls per denoising step, and a feedback VAE decode per step. The sampler records preparation, generation, and feedback-decoding time separately; the table above retains the original runs' costs.

E Evaluation details

E.1 Visual judgments and units

An edit is assessed by the intended local and downstream response, preserved source/body identity, and absence of task-breaking changes elsewhere. Optical full-image and receiver crops are evaluated against source and target; material videos additionally use synchronized temporal strips. Original labels used non-blind automated visual assessment. H1/H2 recovery labels use the authors' subsequent assessment of saved images, not an independent multi-rater study.

Material labels are correct, incorrect, or low-visibility. A configuration is correct in both cameras, incorrect if either fails, or low-visibility when the expected change in the target rendering is too small to distinguish reliably at the displayed scale. Low-visibility describes the observable effect size, not blur or loss of detail introduced by a generated video. All 68 remain in the table. Matching paired labels give 52/0/16 configurations for VECTILE and 6/46/16 for native transfer, or 104/0/32 and 12/92/32 view counts. These are two summaries of the same experiment.

The six native successes are four cold/no-stimulus controls and two cold/blocked controls (r_{00}, r_{03}). Their intended stillness is genuine success. Warm blocked materials can still move because of nonzero state. The older optical model uses method-specific ambiguous labels; these do not become common failures by convention.

E.2 GENERATED-RGB MOTION MEASUREMENTS

Farnebäck flow (Farnebäck, 2003) is measured between consecutive teacher and generated frames after 3×3 Gaussian blur with $\sigma = 0.8$. Parameters are pyramid scale 0.5, three levels, window 19, three iterations, polynomial neighborhood 5, and polynomial sigma 1.1.

Evaluation-only teacher masks locate bodies. A 15-pixel dilation excludes foreground from the background estimate. Per-frame median background flow is subtracted from both teacher and generated flow. Body vectors are averaged in a three-pixel morphological boundary band. For teacher v_t and generated $\widehat{v}_t$, report

$$A = \frac{\sum_t \|\widehat{v}_t\|_2}{\max(\sum_t \|v_t\|_2, 10^{-6})}, \qquad C = \frac{\sum_t \widehat{v}_t^\top v_t}{\max(\|\widehat{v}\|_2 \|v\|_2, 10^{-9})}. \tag{14}$$

Readout requires at least 20 visible pixels per body in every sampled frame. A separate stationarity check uses teacher vertex change below 0.001 times body length. These rules define motion diagnostics, not full-edit visual labels. They measure generated RGB, not the predicted mesh or its VACE controls.

Table 12: Generated-RGB motion diagnostics. Medians over 124 valid body-view records; these quantities do not measure identity or all aspects of editing.

Method	Valid body-views	Amplitude ratio	Direction cosine
Native VIPER-style/Wan	124	0.3054	0.1433
VECTILE	124	1.0854	0.9078

E.3 OBSERVATION, INSTALLATION, AND OUTPUT CAMERA

Reference-camera changes alter the observations used for extraction; installation changes the frame in which a material responds; output-camera changes reproject the completed target trajectory. The installation/observation experiment fixes a target and checkpoint, varies six observations, and repeats for two installations. Small differences between generated images alone can also indicate a consistently wrong static solution. For output-camera tests, codes and initial target state remain fixed. TileFlow computes one trajectory, then each camera receives its own projected controls and synthesis. Correct response in both cameras is evaluated, not exact pixel correspondence between diffusion outputs. Time samples are synchronized; no sequence is warped to align motions.

Table 13: Installation–observation crossing. Body-RGB differences describe appearance variation, not correctness.

Method	Installation	Success/observations	Mean RGB difference
Native VIPER-style	Axis 0°	0/6	0.0032
Native VIPER-style	Axis 90°	0/6	0.0034
VECTILE	Axis 0°	6/6	0.0039
VECTILE	Axis 90°	6/6	0.0040

F COMPOSITION AND ADDITIONAL MATERIAL COMPARISONS

F.1 INDEPENDENTLY INSTALLED RESPONSES

The 32-video multi-response/joint set contains 26 correct and six low-visibility videos, including all four views of $r_{00} \to r_{03}$ in two layouts. The native single-reference interface cannot independently install two different references, so those operations are marked not expressible rather than assigned failure. The multi-slot, target-conditioned extension does express them and is reported separately (Table 14).

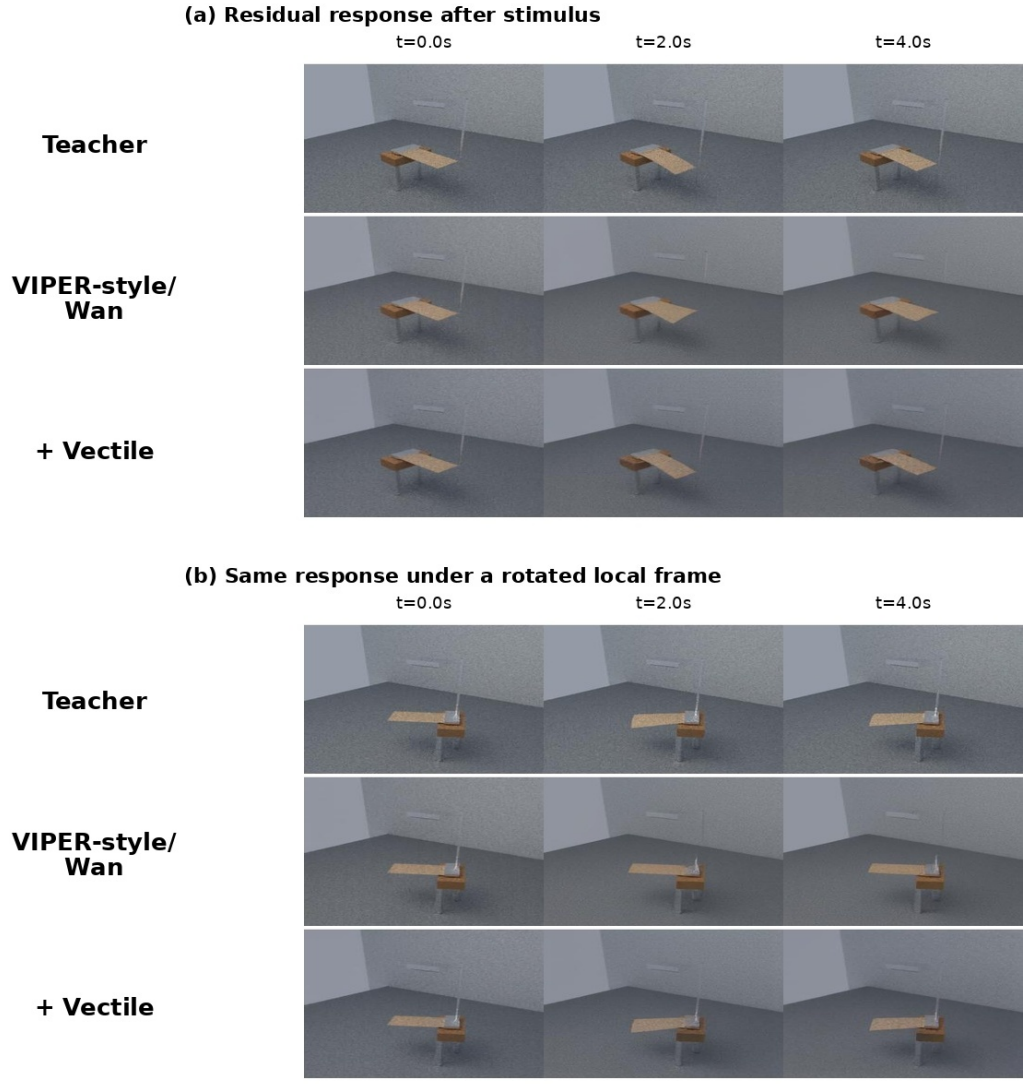

Figure 6: Target history and installation. A warm target continues bending after heat removal; changing the material axis rotates deformation while retaining its response identity. These saved video comparisons preserve the original teacher and generated frames. Input differences are specified in Table 1.

Table 14: Multi-slot, target-conditioned VIPER comparison. The baseline retains Qwen queries, a connector, independent slots, and target conditions. Its training controls also fail, so these are finite-training results, not equally fitted-component comparisons.

Setting	Extended VIPER	VECTILE
Unseen ordered pair, seen world (2 layouts, 2 views)	0/4	4/4
Unmodified training controls (2 configurations, 2 views)	0/4	4/4
Post-hoc factorial (8 configurations, 2 views)	0/16	16/16

This auxiliary checkpoint predates the native-primary comparison and has a distinct training and input history. Extra target information enters its physics-query/connector path, not an executable response branch. It is not pooled with Table 2.

F.2 RESPONSE/SOURCE/INSTALLATION FACTORIAL

The rear response stays at r_{03} while front response, source position, and rear installation are crossed (Table 15). Each row recomputes the target from its edited initial world. Four configurations contain the held pairing $(00, 03)$ and four contain $(02, 03)$; each has two views. This post-hoc seen-world factorial differs from the original rear-r_{01} joint set and does not enlarge the primary denominator.

Table 15: Local reference replacement under fixed target conditions. Rear response r_{03} stays fixed. Within each adjacent pair, source S and rear installation I are fixed; only the front response changes. Fractions give successful generated videos / evaluated views (two per configuration).

Edit	Front	Source	Rear install	VECTILE	Extended VIPER
Original target	r_{00}	S_0	I_0	2/2	0/2
Replace response	r_{02}	S_0	I_0	2/2	0/2
Move source	r_{00}	S_1	I_0	2/2	0/2
Response + source	r_{02}	S_1	I_0	2/2	0/2
Reinstall receiver	r_{00}	S_0	I_1	2/2	0/2
Response + install	r_{02}	S_0	I_1	2/2	0/2
Source + install	r_{00}	S_1	I_1	2/2	0/2
All three	r_{02}	S_1	I_1	2/2	0/2

The paired rows separate response replacement from a change of source or installation. Across pairs, the target changes what the fixed rear response receives; within each pair, only the front response is replaced. These comparisons give the local editing operation and its coupled consequence distinct roles, rather than assigning one precomputed image difference to each response. Figures 7–8 preserve synchronized views of the original outputs.

G ROLLOUT INTERVENTIONS AND INPUT CONSTRUCTION

G.1 STATE AND GEOMETRY

Two interventions use a fixed trained material response. Zeroing supplied temperature, activation, and conduction retains the authored stimulus but removes both queried and carried state. Freezing exposure geometry evaluates incoming heat against initial body positions while response/state updates continue. Table 16 uses selected numerical-world subsets, not the 68-configuration generated-video set. The state intervention is not an isolated ablation of one MLP input channel, because it also changes state carried by the update equations. Their interpretation is that current state and renewed geometry matter to the specified target process.

G.2 LEGAL INITIAL IMAGES AND SITE COVERAGE

In 24 of 314 inspected configurations, the first image showed the original source/installation despite an edit that applied before the supervised first frame. Rendering the edited initial world removes that

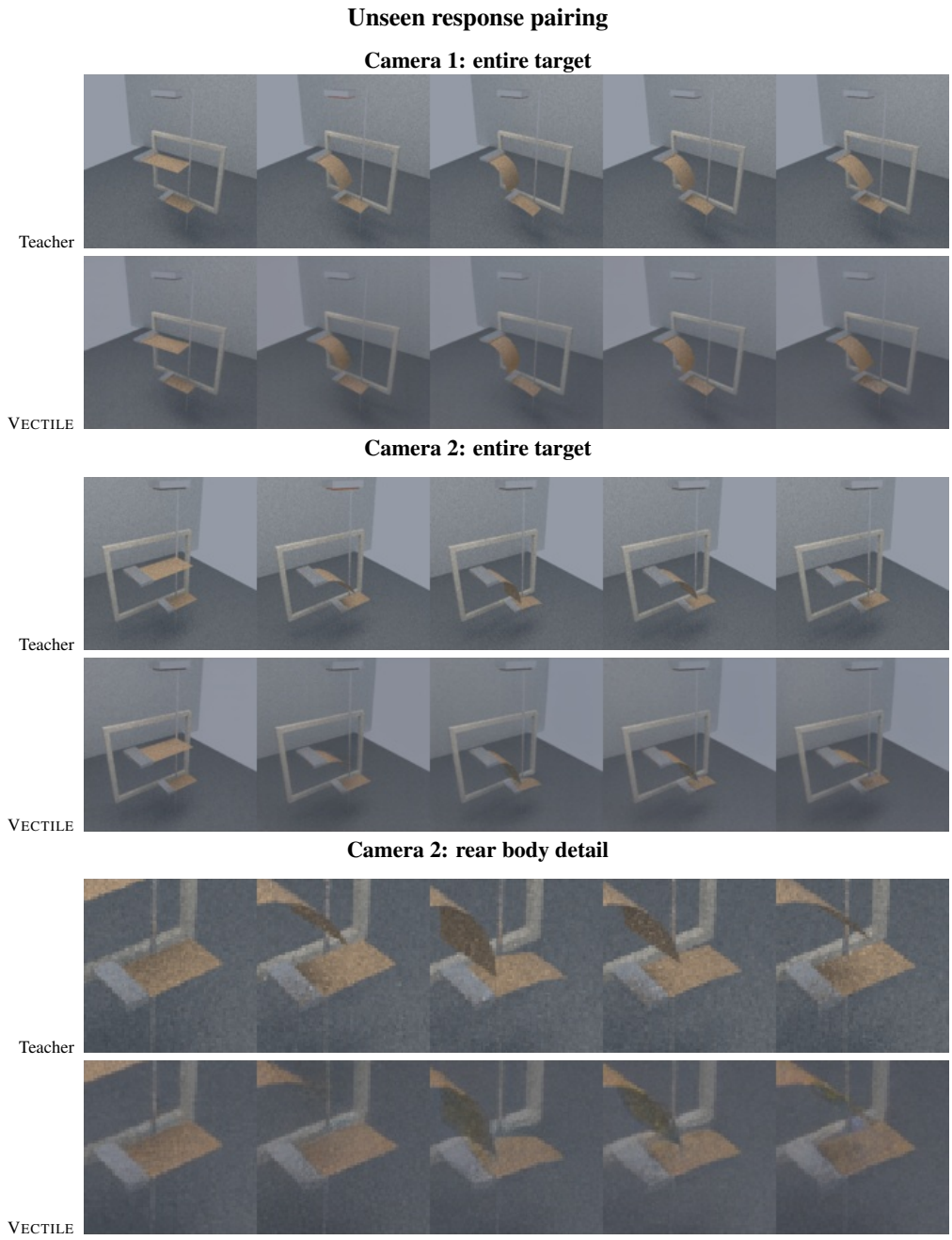

Figure 7: Unseen response pairing. An unseen ordered pairing observed from two cameras. Each teacher strip is followed by the corresponding generated-video strip. The lower pair enlarges the rear body in camera 2. All five columns use the same synchronized times.

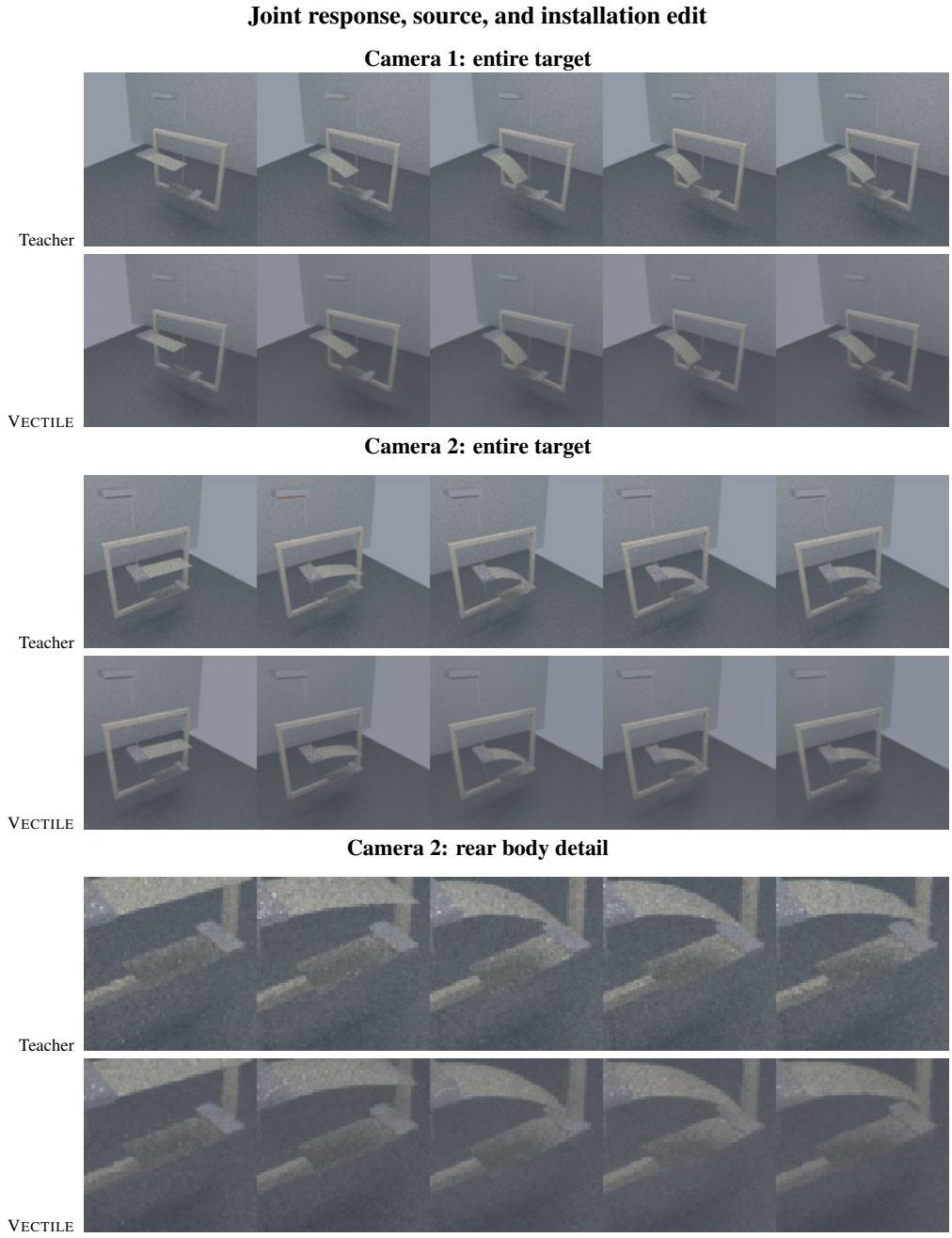

Figure 8: Joint response, source, and installation edit. The front response changes to r_{02} while the rear retains r_{03}; source and rear installation also change. Rows preserve synchronized teacher/generated views and a rear-body detail. This is the additional factorial setting, not the primary held-out count.

Table 16: State and geometry interventions. No model is retrained; subsets and units differ from the final video evaluation.

Intervention	Selected subset	Full	Modified
Zero temperature, activation, conduction	State-sensitive validation	12/12	4/12
Freeze exposure geometry	Multi-body/joint validation	14/14	11/14

conflict for both methods using only initial geometry and authored edits. This is an input-consistency repair, not target-future supervision.

The early optical representation reused a static source-derived 2.5D grid across endpoints, so geometry-changing interventions could fail to change receiving relations. The revised input construction rebuilds endpoint geometry and visibility from the authored edit. Coarse image-grid sampling could also fail to allocate a site to a visible edited surface. Binding cannot address a surface that the discretization omits. The revised construction ensures coverage before response assignment. This observation motivates explicit installation sites; it is not a quantitative learned-model ablation.

G.3 Removal of predicted video controls

A same-checkpoint, same-seed scene comparison removes the predicted VACE control stream while retaining the trained generator. The saved five-case result is 5/5 correct with controls and 0/5 without. Control-off outputs also contain background/object artifacts. The intervention establishes dependence on the trained control distribution; it does not isolate accurate physical content from conditioning-distribution shift, and is not a substitute for response replacement or a matched baseline.

H Primary optical results and the learned-graph alternative

H.1 A learned-graph alternative and the choice of explicit transport

The main paper uses the explicit local-response/typed-transport architecture, designated H2 in the experimental record. We also evaluated H1, a learned-graph alternative with the same spatial-goal and synthesis-control interface.

H1 initializes 64-dimensional site states from geometry and five-condition priors. At each of four rounds, it queries reference tokens, forms sender/receiver messages with learned edge features and visibility gates, and updates its states with a GRU and layer normalization. Local response, propagated influence, and accumulated memory are therefore mixed within the recurrent hidden state. The reference bank is fixed, but its state-dependent retrieval participates in this joint update.

We prefer the explicit architecture because its intermediate variables have distinct roles: a fixed installed code determines P_r, target geometry determines T_W, and the propagated outgoing field becomes the next incident field. This yields a direct correspondence between the response–target factorization and the computation. The learned graph is more flexible in how it combines these quantities, but provides a less transparent separation of response identity, routing, and state. Our choice is architectural clarity, not a claim of superior measured accuracy over H1.

Both implementations retain inherited field conditioning, spatial-goal supervision, and depth-wise control; they recover the same seven tasks below. Their comparison does not isolate typed transport from all other changes to response retrieval and state update. Relative to the earlier optical system, goal prediction, feedback, and training also change, so the image recoveries alone do not isolate recurrence.

H.2 Seven recovered baseline failures

Both variants use the fixed 193 optical inputs and saved baseline images. Subsequent visual assessment examines 69 cases previously labeled unsuccessful for LightMover-style/Wan. H2 recovers seven distinct tasks: one color change and six joint moves (Table 17). H1 recovers the same seven. The criterion accepts minor residual appearance differences when the requested edit, source identity, and scene remain usable; it does not require pixel equality.

Table 17: Seven recovered LightMover-style failures. The typed H2 column is the primary optical implementation; H1 is retained as a learned-graph alternative. Baseline images are unchanged.

Scene	Edit	LM-style	Typed H2	Graph H1
Kitchen q17	Receiver recoloring	Incorrect	Correct	Correct
Bedroom q00	Joint move	Incorrect	Correct	Correct
Bedroom q02	Joint move	Incorrect	Correct	Correct
Bedroom q16	Joint move	Incorrect	Correct	Correct
Living room 2 q00	Joint move	Incorrect	Correct	Correct
Living room 2 q02	Joint move	Incorrect	Correct	Correct
Living room 2 q16	Joint move	Incorrect	Correct	Correct

Kitchen q17 changes both tall receiver panels. The baseline changes the left panel while leaving the right almost white. Both hierarchical outputs restore the missing blue response. Table 18 uses the same right-panel support for every output; these are descriptive measurements, not a replacement success threshold. For joint moves, improvements are clearest in emitter silhouette, target position, and front

Table 18: Right-receiver change in kitchen q17. RGB is normalized to $[0, 1]$. Absolute change is measured from the source; cosine compares its direction with the teacher change.

Image	Mean absolute RGB change	Change-direction cosine
Target	0.1827	1.000
LightMover-style/Wan	0.0087	−0.129
VECTILE typed (H2)	0.1526	0.975
Learned-graph alternative (H1)	0.1462	0.977

installation. Fine lighting differs from the render. Figures 9–10 preserve all seven comparisons without changing image color or exposure. The seven recoveries are not an aggregate H2 success count over 193, and no total adds them to an older checkpoint's successes. The shared recovered subset also does not identify whether gains come from typed transport, a spatial goal, the controller, or additional training.

H.3 ORIGINAL OPTICAL CHECKPOINT AND ADDITIONAL OPERATIONS

Table 19: Complete original optical comparison. This predates H1/H2 and is not their aggregate evaluation. Ambiguous outputs lack reliable binary visual judgments.

Method	Images	Correct	Incorrect	Ambiguous
LightMover-style/Wan	193	46	69	78
Original VECTILE optical	193	42	71	80

Ordinary color edits were often handled by both original methods; source movement and precise receiver response were harder. In six additional original-checkpoint cases, zeroing only the field projection yields 0/6 success both with and without projection and small pixel differences. The six held-out fixed-response movement cases also have no successful RGB edits for either original method. These controls motivate investigating the generation interface, but do not locate every error in either the interaction field or its synthesis.

The separate 100-image original-checkpoint set includes 20 base controls, reference replacements, installation changes, source moves, and joint edits. Native LightMover-style can express the 20 base controls and 12 source moves. Unsupported operations remain dashes (Table 20), not artificial failures. The original source-only real cases 187, 183, 185, and 200 lack authored worlds. Both local original models fail all four; that VECTILE executor is inactive. Dominant differences concern lamp identity and source/receiver synthesis. These outputs are not H1/H2 results and are not the earlier GT-supervised recurrent Stage-3B replay, which has a different primary construction, decoder, and training history.

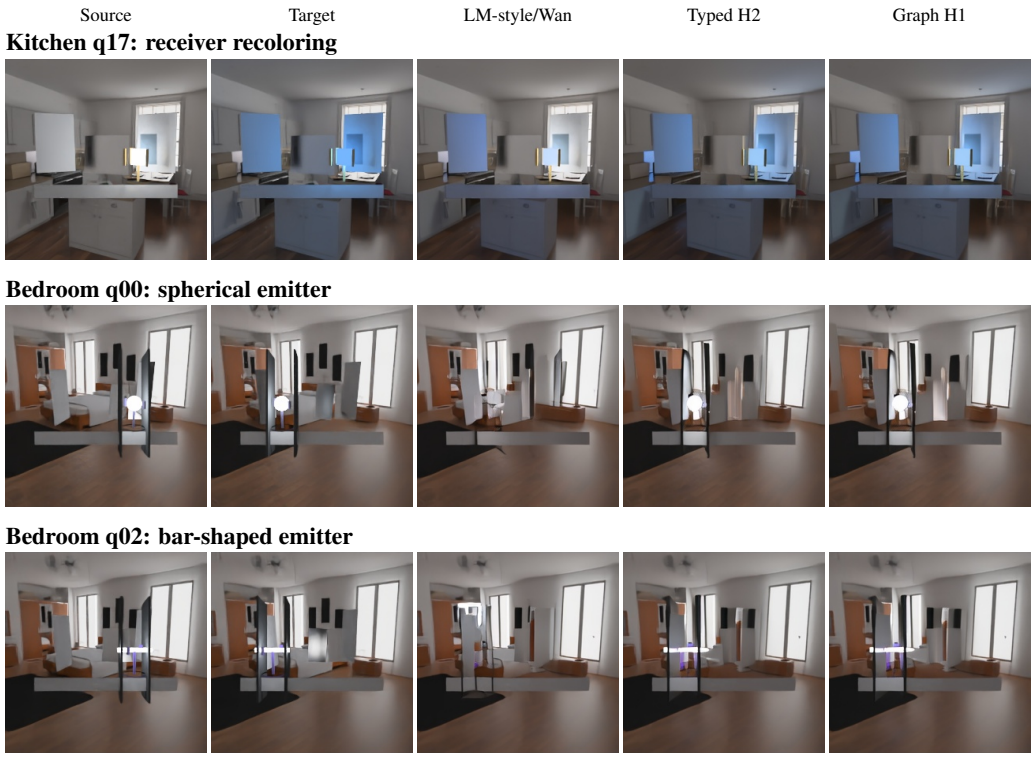

Figure 9: All seven optical recoveries (1/2). Source, target, and unchanged baseline are followed by the primary typed implementation (H2) and the learned-graph alternative (H1). Original image streams and color/exposure are retained.

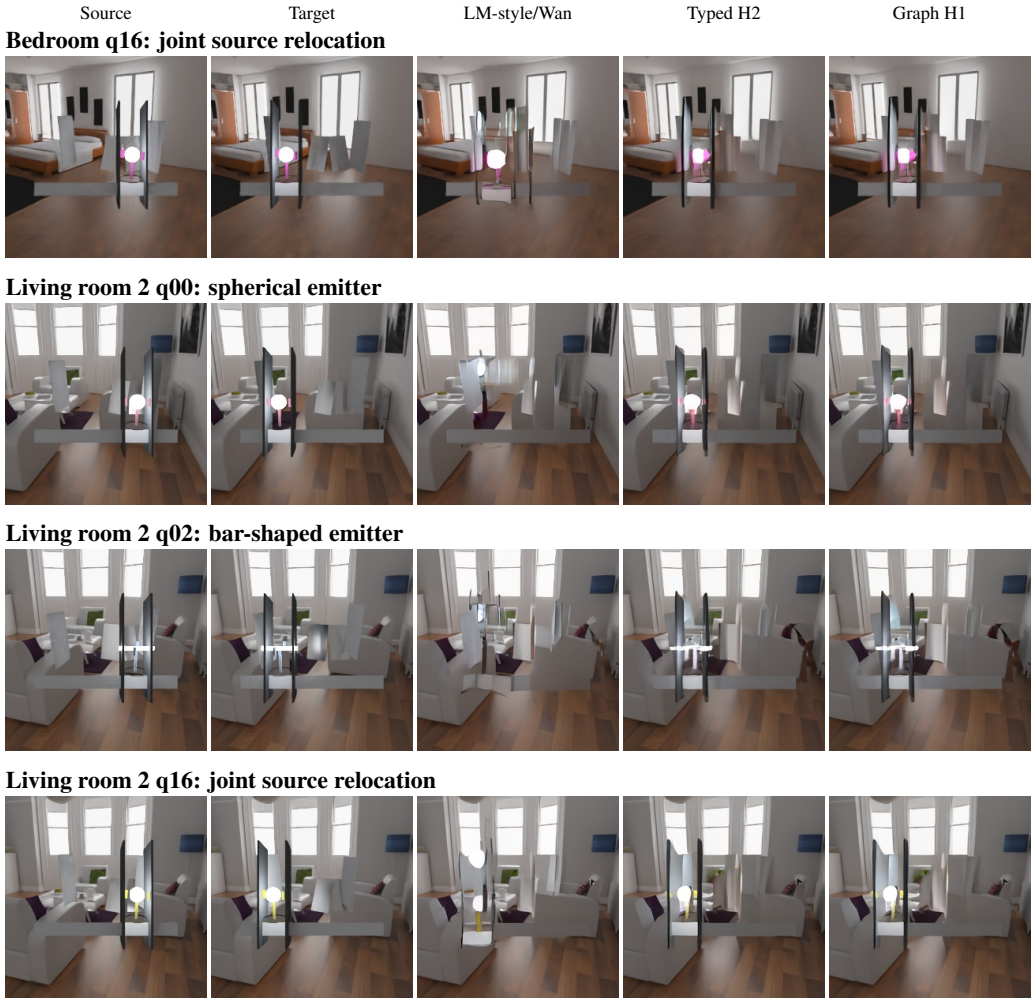

Figure 10: All seven optical recoveries (2/2). Source, target, and unchanged baseline are followed by the primary typed implementation (H2) and the learned-graph alternative (H1). Original image streams and color/exposure are retained.

Table 20: Additional original-checkpoint operations. C/E/A: correct/erroneous/ambiguous. The correct original VECTILE examples are unchanged base controls.

Operation	VECTILE images	VECTILE C/E/A	Native images	Native C/E/A
Base control	20	20/0/0	20	20/0/0
Reference replacement	20	0/9/11	—	—
Receiver installation	20	0/20/0	—	—
Source movement	20	0/20/0	12	0/12/0
Joint edit	20	0/20/0	—	—

I COMPUTATIONAL ENVIRONMENT AND FULL SCENE OUTCOMES

I.1 SOFTWARE AND PRETRAINED BASE MODELS

The recorded local environment is Python 3.13.12 on Windows 11, PyTorch 2.11.0+cu130, NumPy 2.4.4, Diffusers 0.40.0, Transformers 5.5.0, Accelerate 1.15.0, Safetensors 0.7.0, Pillow 12.2.0, and SciPy 1.17.1, on an RTX 5070 Ti. CPU Mitsuba and the shell solver use separate compatible environments. The host identifier is Wan-AI/Wan2.1-VACE-1.3B-diffusers, revision `ec4d2cb062b548996b179d493fdd05340de702a1`; the MLLM is Qwen/Qwen3-VL-4B-Instruct, revision `ebb281ec70b05090aa6165b016eac8ec08e71b17`.

I.2 PER-CONFIGURATION PRIMARY SCENE OUTCOMES

Table 21 lists all 68 saved primary configuration judgments. The common prefix `world_933000_` is omitted. S/F/L denote success/failure/low-visibility; both views of each listed configuration have the same recorded label. Native denotes the native VIPER-style/Wan model.

Table 21: All primary configurations. The fixed prefix is omitted; each row has two views.

Configuration	Native	VECTILE
`context_response_00_cold_blocked_stimulus`	S	S
`context_response_00_cold_full_stimulus`	F	S
`context_response_00_cold_long_pulse`	F	S
`context_response_00_cold_no_stimulus`	S	S
`context_response_00_cold_weak_stimulus`	F	S
`context_response_00_warm_after_stimulus`	F	S
`context_response_00_warm_blocked_stimulus`	F	S
`context_response_00_warm_short_pulse`	F	S
`context_response_01_cold_full_stimulus`	F	S
`context_response_01_cold_long_pulse`	F	S
`context_response_01_cold_no_stimulus`	S	S
`context_response_01_cold_weak_stimulus`	L	L
`context_response_01_warm_after_stimulus`	L	L
`context_response_01_warm_short_pulse`	F	S
`context_response_02_cold_full_stimulus`	F	S
`context_response_02_cold_long_pulse`	F	S
`context_response_02_cold_no_stimulus`	S	S
`context_response_02_cold_weak_stimulus`	F	S
`context_response_02_warm_after_stimulus`	L	L
`context_response_02_warm_short_pulse`	F	S
`context_response_03_cold_blocked_stimulus`	S	S
`context_response_03_cold_full_stimulus`	F	S
`context_response_03_cold_long_pulse`	F	S
`context_response_03_cold_no_stimulus`	S	S
`context_response_03_cold_weak_stimulus`	L	L
`context_response_03_warm_after_stimulus`	F	S
`context_response_03_warm_blocked_stimulus`	F	S
`context_response_03_warm_short_pulse`	F	S
`frame_response_00_left_axis0_yaw+30`	F	S

Continued on next page.

Table 21 continued.

Configuration	Native	VECTILE
frame_response_00_left_axis0_yaw-30	F	S
frame_response_00_left_axis90_yaw+30	F	S
frame_response_00_left_axis90_yaw-30	F	S
frame_response_00_right_axis0_yaw+30	F	S
frame_response_00_right_axis0_yaw-30	F	S
frame_response_00_right_axis90_yaw+30	F	S
frame_response_00_right_axis90_yaw-30	F	S
frame_response_01_left_axis0_yaw+30	F	S
frame_response_01_left_axis0_yaw-30	F	S
frame_response_01_left_axis90_yaw+30	L	L
frame_response_01_left_axis90_yaw-30	L	L
frame_response_01_right_axis0_yaw+30	F	S
frame_response_01_right_axis0_yaw-30	F	S
frame_response_01_right_axis90_yaw+30	L	L
frame_response_01_right_axis90_yaw-30	L	L
frame_response_02_left_axis0_yaw+30	F	S
frame_response_02_left_axis0_yaw-30	F	S
frame_response_02_left_axis90_yaw+30	L	L
frame_response_02_left_axis90_yaw-30	L	L
frame_response_02_right_axis0_yaw+30	F	S
frame_response_02_right_axis0_yaw-30	F	S
frame_response_02_right_axis90_yaw+30	L	L
frame_response_02_right_axis90_yaw-30	L	L
frame_response_03_left_axis0_yaw+30	F	S
frame_response_03_left_axis0_yaw-30	F	S
frame_response_03_left_axis90_yaw+30	F	S
frame_response_03_left_axis90_yaw-30	F	S
frame_response_03_right_axis0_yaw+30	F	S
frame_response_03_right_axis0_yaw-30	F	S
frame_response_03_right_axis90_yaw+30	F	S
frame_response_03_right_axis90_yaw-30	F	S
reuse_response_00_mount0_source0_state0	F	S
reuse_response_00_mount0_source0_state1	F	S
reuse_response_00_mount0_source1_state0	F	S
reuse_response_00_mount0_source1_state1	F	S
reuse_response_00_mount1_source0_state0	L	L
reuse_response_00_mount1_source0_state1	L	L
reuse_response_00_mount1_source1_state0	L	L
reuse_response_00_mount1_source1_state1	L	L