CAER: Causal Action Effect Reweighting for World Model Training

Jianjie Fang[1,*], Xvyuan Liu[1,*], Ziyou Wang[1], Rongze Tang[3], Zhaolu Wang[1], Zhuohang Li[1], Xin Zhang[2], Haisheng Su[2], Chen Gao[1,†], Wei Wu[2], Xinlei Chen[1,†], Yong Li[1,†]
[1]Tsinghua University [2]Manifold AI [3]University of Science and Technology of China
chgao96@gmail.com, chen.xinlei@sz.tsinghua.edu.cn,
liyong07@tsinghua.edu.cn
[*]**Equal contribution.** [†]**Corresponding authors.**

Abstract. World models are becoming core infrastructure for embodied intelligence, with action-conditioned video generation providing controllable predictions of how scenes evolve after agent interventions. Yet existing models are commonly trained with space–time-uniform mean squared error, allowing abundant background tokens to dominate the gradient while sparse interaction dynamics remain under-optimized; such uniform fitting rewards reconstructing appearance rather than learning how actions change the world. We introduce **C**ausal **A**ction **E**ffect **R**eweighting **(CAER)**, a general training paradigm that redistributes supervision toward the tokens whose predicted future is causally affected by the action. CAER contrasts the model's own predictions with and without action conditioning to localize these tokens online, then normalizes the resulting effect map into a weight that preserves the total coefficient mass and changes only where it is spent. This online signal requires no external annotations or offline preprocessing, avoids additional data-processing time, and scales naturally with model and dataset size. Experiments across heterogeneous action-conditioned world-model tasks show that CAER converges to better solutions than uniform MSE training, with consistent improvements in the physical consistency, controllability, and visual quality of generated videos.

Date: August 31, 2026
Webpage: https://manifoldai-research.github.io/CAER/

1 Introduction

World models equip embodied agents with the ability to look beyond the current observation and anticipate how the environment may evolve under different behaviors. By turning learned dynamics into simulated experience, they support planning, evaluation, and policy learning before actions are executed in the physical world. Advances in video diffusion and flow-matching backbones [1–3] have made visually realistic world simulation increasingly feasible, supporting robotics, interactive environments, and policy development beyond what can be safely or economically collected in the physical world [4–9]. Within this paradigm, action-conditioned video generation provides a direct route to controllable prediction by modeling how visual scenes evolve in response to agent interventions. For an embodied agent, however, plausible appearance alone is insufficient. A useful world model must capture the physical consequences of interaction: where contact occurs, when an object begins to move, how forces propagate through the scene, and how the environment responds over time.

Despite this progress, current world models are still predominantly optimized with mean squared error averaged uniformly over space and time. Most video tokens describe static backgrounds, slowly

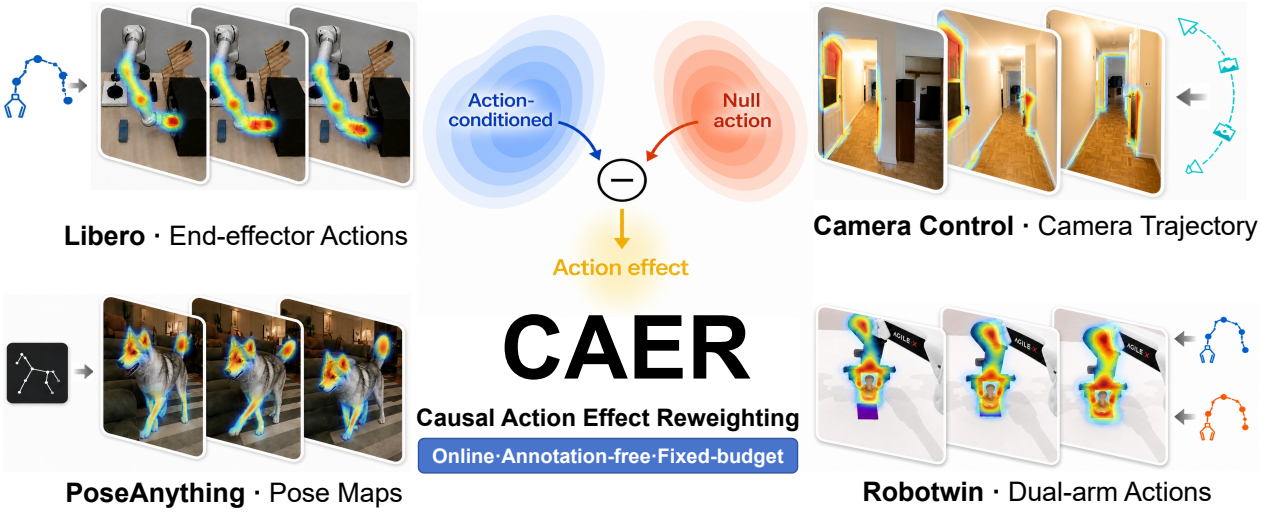

Figure 1: CAER overview. CAER identifies and reweights action-responsive tokens online without external annotations. Examples span end-effector trajectories, camera trajectories, dual-arm actions, and pose maps.

varying context, or regions unrelated to the current interaction, whereas the decisive prediction errors are concentrated around brief contact events, manipulated objects, and actor motion. Uniform averaging therefore allows abundant, easily reconstructed tokens to dominate the gradient and dilute supervision for sparse physical interactions. This mismatch is especially damaging in embodied settings: a rollout may look realistic on average while placing contact at the wrong time, moving an object in the wrong direction, or failing to respond to the commanded action. Such failures are not merely visual defects; they indicate that the model has not learned how actions causally alter the world.

One possible remedy is to annotate actors, objects, motion, and contact regions explicitly. Segmentation models such as SAM [10] and optical-flow estimators such as RAFT [11] can track these regions and provide handcrafted loss weights. This solution, however, is difficult to scale to world-model data: per-frame inference introduces substantial computation, storage, and preprocessing overhead, while the resulting supervision inherits the biases and failure modes of the external models. It also defines relevance through a fixed annotation pipeline rather than through the causal relationship between the action and the predicted future.

In this work, we ask whether the world model itself can provide the missing supervision signal. Our key observation is that additional gradient should be assigned not to every visually salient region, but to tokens whose predicted future is sensitive to the action. If removing the action changes the prediction at a token, that token captures an action-dependent consequence that the model must learn, whereas a token that is unaffected by the action carries no action-conditional risk however much it moves. This turns loss allocation into an online causal action-response problem. The model needs no external definition of the actor, manipulated object, or contact interface, because it identifies relevant interactions through its own conditional response. This principle is action-agnostic: it applies to end-effector poses, joint commands, spatial action maps, and camera trajectories alike, whenever a null-action condition can be learned, which aligns it with the broader goal of grounding world models in the consequences of intervention rather than only in perceptual likelihood [12, 13].

To examine whether self-computed causal action effects can serve as a scalable reweighting signal for world model training, we investigate three research questions.

- **RQ1: Does action-effect reweighting improve interaction-region prediction quality?** Since

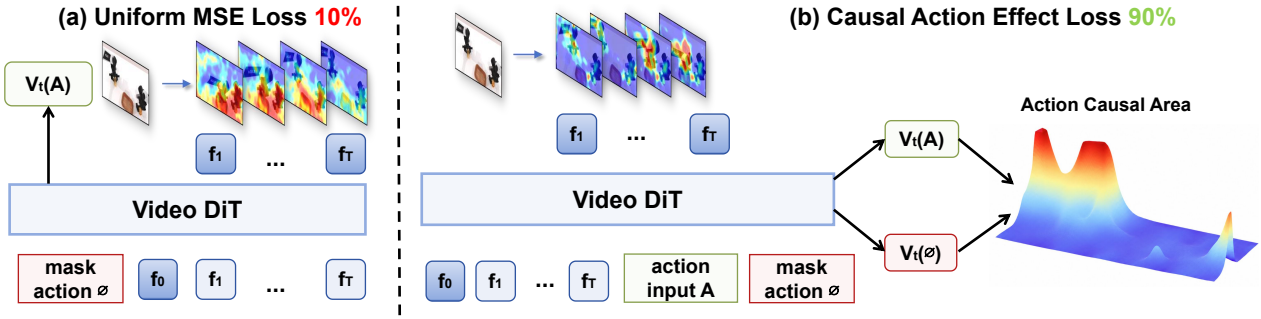

Figure 2: CAER framework overview. At a fixed intermediate noise level and under shared noise, CAER contrasts the model's response under the real action and a learned null action to localize action-sensitive tokens, then normalizes the map sample-wise to unit mean. The objective redistributes a fixed coefficient mass toward causally responsive regions, while action dropout keeps every token stochastically supervised.

token-uniform losses are the default for generative world models, it remains unclear whether reweighting toward unresolved action effects improves the regions that matter most—contact events, manipulated objects, and actor motion—while preserving background fidelity and overall visual quality. Answering this question establishes whether CAER is an effective replacement for uniform reconstruction loss in action-conditioned settings.

- **RQ2: How do the key design choices contribute to the effect map quality and training stability?** CAER introduces two training choices that govern the quality of the online counterfactual signal: the action-dropout rate that defines the null branch, and the fixed noise level at which the effect is read. We vary each in isolation to understand its contribution to convergence speed, loss stability, and final generation performance.
- **RQ3: Does the self-computed signal sharpen as the model improves?** A central motivation of self-computed reweighting is that the supervision signal improves together with the model, forming a training loop that needs no external annotation. We test whether the benefit of the self-computed signal grows with training, overtaking uniform averaging once the causal interaction is learned, and whether the resulting gains appear in task-level success rather than in appearance metrics alone.

Building on this observation, we introduce **CAER** (Fig. 1), a causal action effect reweighting paradigm that changes how the conventional MSE-based flow-matching objective allocates supervision. Instead of averaging all space–time tokens uniformly, CAER compares predictions under the observed action and a learned null-action condition at a fixed intermediate noise level, using their velocity-space difference to identify where the predicted future responds to the action. Action dropout keeps this counterfactual query in distribution. The action response is then normalized sample-wise into a weight that preserves the total coefficient mass and changes only where it is spent. Because the coefficient mass is fixed, the comparison with uniform averaging becomes a pure allocation question: to first order in the learning rate, focused weighting reduces interaction-region risk faster than uniform averaging whenever the weight is positively correlated with token utility, and it still descends at points where uniform training stalls because sparse interaction gradients are cancelled by dominant background gradients. Because the complete signal is computed online by the world model, CAER requires no segmentation, tracking, optical flow, or contact annotation and applies to both spatial and numerical action interfaces.

Experiments show that this causal redistribution improves interaction modeling and visual generation together. Because the weight is normalized rather than added, CAER can be enabled from the first

optimization step without retuning the loss scale, and its focusing signal is refreshed at every step instead of being fixed in advance. Across action-conditioned tasks with distinct control modalities, it converges to better solutions than the matched uniform-MSE baseline, with stronger visual quality and physical consistency. As training proceeds, the focus contracts onto causal interaction regions and those dynamics are predicted with progressively greater accuracy. The resulting world model therefore captures not only sharper appearance, but also when interactions occur, how objects respond to interventions, and how agents affect the surrounding world.

2 Methodology

CAER allocates gradient to locations where the predicted future is causally sensitive to the action, a signal read online from the world model itself without segmentation, tracking, optical flow, or contact labels. Motion cues identify *what moves*, whereas control requires *what the action governs*: camera shake and exogenous motion are salient under the former yet carry no action-conditional risk. Crucially, the weight redistributes a fixed coefficient mass instead of enlarging it, which turns the comparison with uniform training into a pure allocation question and lets us characterize, at equal mass, when the reallocation provably wins (Fig. 2).

2.1 Problem Formulation

Let the backbone be a pretrained video generator and $z_0 \in \mathbb{R}^{B\times C\times T\times h\times w}$ the video latent, where $t = 0$ denotes the reference frame and $t = 1,\dots,T-1$ the supervised future. We use the linear flow-matching path

$$\begin{aligned} z_\tau &= (1-\tau)z_0 + \tau\epsilon, \\ v_{\text{tgt}} &= \epsilon - z_0, \qquad \epsilon \sim \mathcal{N}(0, I), \end{aligned} \tag{1}$$

where τ follows the backbone's timestep distribution. The model predicts $v_\theta(z_\tau; c_A)$ conditioned on $c_A = \{\mathbf{o}_0, \mathbf{p}, \mathcal{A}\}$, containing the reference observation, language instruction, and action condition. Uniform flow matching averages token errors, allocating gradient by token count rather than by downstream utility: static background tokens dominate the coefficient mass although generation failures concentrate on sparse actor–object interactions.

2.2 Action Conditioning and Effect Estimation

Action conditioning. A frame-aligned action sequence $A = [a_1,\dots,a_K]$, with $a_k \in \mathbb{R}^{d_a}$, is mapped to a backbone-compatible condition $\mathcal{A} = \phi(A)$. The encoder ϕ follows the backbone's control interface, either spatial, where frame-wise action maps are VAE-encoded onto the video latent grid, or numerical, where action vectors modulate each Transformer block (Table 1). New pathways are zero-initialized where possible, preserving the pretrained image-to-video function at initialization.

Null action. Reweighting only requires that removing $\mathcal{A}$ produces a well-defined prediction $v_\theta(z_\tau; c_\emptyset)$, where $c_\emptyset = \{\mathbf{o}_0, \mathbf{p}, \emptyset\}$. We realize $\emptyset$ by zeroing or masking the injected action features, whichever form the interface exposes. Because an action-free input would otherwise be out of distribution, we independently disable action injection with probability $p_{\text{drop}} = 0.10$ and predict the same target from $c_\emptyset$. Similar to classifier-free guidance [14], this trains a usable null branch. Here it also keeps both CAER queries in distribution and provides the recall floor in Eq. (5).

Action effect map. We evaluate a controlled counterfactual: under the same history and noisy latent, where does the predicted velocity change when the true action is removed? Using shared noise ϵ', we build a latent at a fixed intermediate time $\tau_S = 0.50$ by $z_{\tau_S} = (1-\tau_S)z_0 + \tau_S\epsilon'$, and reduce the two

responses along the channel dimension:

$$S = \left\| v_\theta(z_{\tau_S}; c_A) - v_\theta(z_{\tau_S}; c_\emptyset) \right\|_2 \tag{2}$$

with $S \in \mathbb{R}^{B\times 1\times T\times h\times w}$. Sharing ϵ' removes sampling variation, while fixing τ_S prevents the scale and spatial structure of S from drifting with the random training timestep. The two evaluations are extra forwards under no-gradient mode, run alongside the single gradient-carrying forward at the sampled τ that produces the loss. On the linear path $\hat{z}_0 = z_{\tau_S} - \tau_S v_\theta$, so $S = \tau_S^{-1}\|\delta_\theta\|_2$ for the action-induced increment between the model's two clean-state predictions, $\delta_\theta = \hat{z}_0(z_{\tau_S}; c_A) - \hat{z}_0(z_{\tau_S}; c_\emptyset)$, which is the model's current estimate of

$$\delta = \mathbb{E}\left[z_0 \mid z_{\tau_S}, c_A\right] - \mathbb{E}\left[z_0 \mid z_{\tau_S}, c_\emptyset\right]. \tag{3}$$

Thus S measures how strongly the model's predicted future depends on the action, which is what separates it from flow or segmentation: a token with negligible δ_i carries little action-conditional risk regardless of its pixel-space motion. Since z_{τ_S} is a noised version of the realized future and hence itself downstream of the action, Eq. (3) is an action-conditional predictive contrast rather than an interventional quantity in the sense of do-calculus; we use *causal* in this operational sense throughout, which suffices because the objective needs only a ranking of tokens by residual action dependence, not an identified causal graph.

2.3 Focused Objective and Optimization Effect

Budget-neutral weight. Let $\Omega_b = \{(t, x) : t = 1, \dots, T-1\}$ be the $|\Omega_b| = (T-1)hw$ future tokens of sample b, and $\mu^{(b)}(F) = |\Omega_b|^{-1}\sum_{(t,x)\in\Omega_b} F_x^{b,t}$ the future-token mean of a scalar map F. Because the magnitude of S changes with training progress and action amplitude, we normalize it independently for each sample:

$$\rho_x^{b,t} = \frac{S_x^{b,t}}{\max\left\{\mu^{(b)}(S), \varepsilon_0\right\}}, \qquad \mu^{(b)}(\rho) = 1, \tag{4}$$

with a small constant $\varepsilon_0 > 0$ for numerical safety. The unit-mean constraint fixes the total coefficient mass to that of uniform training, so reweighting redistributes coefficient mass within each sample without increasing its total mass or shifting mass across samples. The construction of ρ is detached from the backward graph.

Recall floor. A token with zero measured effect would receive no gradient, so a false negative could never be corrected, whereas a false positive only spends part of one update; the error is therefore asymmetric. Action dropout removes this failure mode at no additional cost. On dropped samples we skip Eq. (2) and set $\rho \equiv 1$, supervising every token with unit weight through the null branch, which shares all parameters with the action branch. Averaged over the dropout variable, the coefficient applied to token (t, x) is

$$\begin{aligned} \bar{\rho}_x^{b,t} &= (1 - p_{\text{drop}})\rho_x^{b,t} + p_{\text{drop}}, \\ \bar{\rho}_x^{b,t} &\geq p_{\text{drop}} > 0, \qquad \mu^{(b)}(\bar{\rho}) = 1, \end{aligned} \tag{5}$$

so the effective weight field is strictly positive and no token is permanently excluded from optimization, while the coefficient mass is preserved exactly rather than inflated by a hand-set lower bound; because dropped samples are supervised under $c_\emptyset$, Eq. (5) is the coefficient reaching token (t, x) through the shared parameters. With c_b the condition used in sample b's main forward, and $\rho \equiv 1$ on dropped samples, the token error is

$$\ell_x^{b,t} = \left\| v_\theta(z_\tau; c_b)_x^{b,t} - v_{\text{tgt}x}^{b,t} \right\|_2^2. \tag{6}$$

Since $\sum_{(t,x)\in\Omega_b} \rho_x^{b,t} = |\Omega_b|$, no additional normalization is required:

$$\mathcal{L}_{\text{focus}} = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{|\Omega_b|} \sum_{(t,x)\in\Omega_b} \text{sg}\left(\rho_x^{b,t}\right) \ell_x^{b,t}. \tag{7}$$

Allocation gain at equal coefficient mass. We drop the sample index and write $g_i = \nabla_\theta \ell_i$ for $i \in \Omega$. For a tolerance $\kappa \geq 0$, let $R = \{i : \|\delta_i\| > \kappa\}$ collect the tokens whose predicted future is appreciably action-dependent, and $R^c = \Omega \setminus R$ the background. Define the interaction risk, the region gradients, and the utility of token i as

$$\mathcal{J}_R = \frac{1}{|R|}\sum_{i\in R} \ell_i, \quad G_R = \frac{1}{|\Omega|}\sum_{i\in R} g_i, \quad a_i = \langle \nabla \mathcal{J}_R, g_i \rangle, \tag{8}$$

with G_{R^c} defined analogously, so that $\nabla \mathcal{J}_R = \frac{|\Omega|}{|R|} G_R$ and $G_{\text{uni}} = G_R + G_{R^c}$. Comparing the equal-mass updates $G_{\text{uni}} = |\Omega|^{-1} \sum_i g_i$ and $G_\rho = |\Omega|^{-1} \sum_i \rho_i g_i$ to first order and using $\mu(\rho) = 1$,

$$\begin{aligned} &\mathcal{J}_R(\theta - \eta G_{\text{uni}}) - \mathcal{J}_R(\theta - \eta G_\rho) \\ &\qquad = \eta \ \text{Cov}_i(\rho_i, a_i) + O(\eta^2), \end{aligned} \tag{9}$$

where the covariance is taken over tokens drawn uniformly from Ω. Positive covariance between weight and utility is therefore exactly the condition, to first order in η, under which focused weighting reduces interaction risk faster than uniform weighting at identical coefficient mass.

Equation (2) supplies that correlation. For $i \in R$ the utility decomposes as

$$a_i = \frac{1}{|R|}\left(\|g_i\|_2^2 + \sum_{j\in R\setminus\{i\}} \langle g_j, g_i \rangle \right), \tag{10}$$

whose diagonal term is strictly positive whenever the token is not yet solved, whereas a token in R^c enters $\mathcal{J}_R$ only through parameter sharing and contributes cross terms of no systematic sign. Since $S_i \to \|\delta_i\|/\tau_S$ as the estimate improves, ρ places above-average weight precisely on the tokens that carry this positive diagonal term. Only a positive association between ρ and a is required, not calibrated magnitudes, so the argument tolerates estimation error in S, and it is the association across tokens rather than the sign of any single a_i that matters.

Task setting	Training source	Samples (clips × frames)	Action injection	Evaluation benchmark
LIBERO [15]	LIBERO simulator rollouts	44,166 × 17	Low-dimensional actions → MLP modulation	WorldArena [16]
RoboTwin [17]	RoboTwin 2.0	50,016 × 17	14-D actions → MLP modulation	WorldArena [16]
Camera Control	RealEstate10K [18]	30,016 × 81	Camera trajectory → Control Adapter	iWorld-Bench [9]
PoseAnything [19]	Official training set	51,792 × 17	Pose maps → VAE visual branch	VBench [20]

Table 1: Training and evaluation settings. All experiments use Wan 2.2 5B as the backbone.

Escaping cancellation. Uniform descent halts wherever $G_{\text{uni}} = G_R + G_{R^c} = 0$, which can occur with $G_R \neq 0$: sparse interaction gradients are cancelled by background gradients even though $\nabla \mathcal{J}_R \propto G_R \neq 0$

leaves the interaction risk locally reducible. Let α and β denote the mean weights over R and R^c. Unit mass forces

$$\frac{|R|}{|\Omega|}\alpha + \frac{|R^c|}{|\Omega|}\beta = 1, \tag{11}$$

so a single one-sided input suffices: (A1) the effect map is informative in the weak sense that $\alpha > 1$. Unit mass then forces $\beta < 1$, and $\alpha > \beta$ follows from Eq. (4) rather than from a separate assumption about the background. Treating ρ as constant within each region, at such a point θ_u

$$\begin{aligned} G_\rho(\theta_u) &= \alpha G_R + \beta G_{R^c} = (\alpha - \beta) G_R \neq 0, \\ \left\langle \nabla \mathcal{J}_R, G_\rho(\theta_u) \right\rangle &= (\alpha - \beta)\frac{|\Omega|}{|R|}\|G_R\|_2^2 > 0. \end{aligned} \tag{12}$$

Focused weighting therefore produces a strict first-order decrease of the interaction risk under (A1) alone, with no assumption on the alignment between G_R and G_{R^c}, and the guarantee is quadratic in the very gradient that uniform training cancelled away. Because ρ is recomputed and detached at every step, the result applies to the surrogate optimized at that step.

Shared optimum and variance. After stop-gradient the weights are constants, and averaging over the dropout variable gives $\bar{\rho} \geq p_{\text{drop}} > 0$, so the population objective is a strictly positive weighted sum of token-wise Bregman divergences. Because ρ is built from z_0, it is not measurable with respect to the conditioning alone, so whether reweighting moves the population optimum is a substantive question. Writing $\mathbb{E}_c[\cdot]$ for the expectation given (z_τ, c), the per-token minimizer is available in closed form and its deviation from the uniform optimum $v^* = \mathbb{E}[v_{\text{tgt}} \mid z_\tau, c]$ is exactly one covariance,

$$v_i^\rho = \frac{\mathbb{E}_c\left[\bar{\rho}_i v_{\text{tgt},i}\right]}{\mathbb{E}_c[\bar{\rho}_i]}, \qquad v_i^\rho - v_i^* = \frac{\operatorname{Cov}_c\left(\bar{\rho}_i, v_{\text{tgt},i}\right)}{\mathbb{E}_c[\bar{\rho}_i]}, \tag{13}$$

taken componentwise between the scalar weight and the vector target. Uniform flow matching is recovered whenever the effective weight is uncorrelated with the aleatoric part of the target, and any violation is damped by a denominator bounded below by p_{drop} through Eq. (5). The construction keeps the numerator small by design: both branches share z_{τ_S}, so common structure cancels and only the action-induced difference survives; the channel norm discards its sign, removing direct coupling to a signed residual; and the map is read at a fixed τ_S from an independent draw ϵ', so it is not a function of the noise defining v_{tgt} at τ. A dependence on z_0 persists because z_{τ_S} contains it, and Eq. (13) states precisely how much that can matter; up to this bounded tilt, the points escaped in Eq. (12) are optimization-induced stationary points rather than alternative population optima. The remaining cost is second order, entering through the second moment $\mu(\rho^2) = 1 + \chi^2(\rho \,\|\, \text{unif})$, which bounds the variance inflation of the token-averaged gradient: it stays near one while $\rho \approx 1$ early in training and grows only as the weight concentrates. As θ improves, its counterfactual estimate sharpens and the covariance in Eq. (9) can increase, while Eq. (5) keeps every token receiving gradient, closing a self-consistent loop that needs no external annotation.

3 Experiments

We evaluate CAER against uniform MSE under a matched training protocol across four action-conditioned settings, then analyze key hyperparameters and the evolution of the focus map during training.

Table 2: Uniform MSE versus CAER on iWorld-Bench [9] for camera control. Aggregate plus generation-quality and trajectory-following metrics; memory metrics are omitted. Higher is better; best in **bold**.

		Aggregate	Generation Quality				Trajectory Following	
	Objective	Avg.	Image Quality	Brightness Consistency	Color Temp. Constraint	Sharpness Retention	Motion Smoothness	Trajectory Accuracy
Camera Control	Uniform MSE	0.6412	**0.6898**	0.5513	0.5196	**0.4752**	0.9902	**0.6211**
	CAER	**0.6614**	0.6804	**0.6693**	**0.6627**	0.4150	**0.9934**	0.5474

Table 3: Uniform MSE versus CAER on WorldArena [16] for LIBERO and RoboTwin. EWMScore and all 15 constituent metrics across six dimensions. Higher is better; best within each task in **bold**.

		Aggregate	Visual Quality			Motion Quality			Content Consistency			Physics Adherence		3D Accuracy		Controllability	
	Objective	EWM Score	Image Quality	Aesthetic Quality	JEPA Similarity	Dynamic Degree	Flow Score	Motion Smoothness	Subject Consistency	Background Consistency	Photometric Consistency	Interaction Quality	Trajectory Accuracy	Depth Accuracy	Perspectivity	Instruction Following	Semantic Alignment
LIBERO	Uniform MSE	57.66	0.3655	0.4950	0.5472	0.1383	0.0388	0.4853	0.6050	0.6280	0.6829	**0.5900**	0.8111	**0.9805**	**0.8080**	0.5520	0.9207
	CAER	**61.79**	**0.3694**	**0.5076**	**0.5639**	**0.1623**	**0.0549**	**0.5020**	**0.7624**	**0.7915**	**0.8397**	0.5885	**0.8475**	0.9793	0.8077	**0.5615**	**0.9300**
RoboTwin	Uniform MSE	62.35	0.5200	**0.3895**	**0.8323**	0.4667	0.2738	0.7930	**0.8234**	**0.8967**	**0.1232**	0.6505	**0.2781**	0.8093	0.9071	**0.6929**	**0.8955**
	CAER	**63.13**	**0.5502**	0.3879	0.8187	**0.5268**	**0.3251**	**0.8231**	0.8125	0.8957	0.0986	**0.6620**	0.2610	**0.8212**	**0.9180**	0.6860	0.8825

Table 4: Uniform MSE versus CAER on VBench [20] for PoseAnything. Results on six fine-grained quality dimensions. Higher is better; best in **bold**.

	Objective	Aggregate	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
PoseAnything	Uniform MSE	0.7422	0.8460	**0.9273**	**0.9686**	0.74	**0.4348**	0.5366
	CAER	**0.7746**	**0.8828**	0.9202	0.9608	**0.86**	0.4269	**0.5971**

3.1 Experimental Settings

Data and control interfaces. All four settings use Wan 2.2 5B [1]; their data, control interfaces, and evaluation benchmarks are summarized in Table 1. Robot actions are temporally aligned with the video clips and injected through MLP-based DiT modulation. Camera trajectories are projected into patch-level control features, whereas frame-wise pose maps are VAE-encoded to align with the video latent grid.

Matched training protocol. Within each setting, uniform MSE and CAER share the initialization, action pathway, training samples and order, optimizer, schedule, batch size, resolution, and training steps; only the objective changes. All experiments use eight NVIDIA H20 GPUs. LIBERO, RoboTwin, and camera control are evaluated after 5,000 optimization steps, while PoseAnything is evaluated after 1,000 steps. Unless otherwise stated, CAER uses $p_{\text{drop}} = 10\%$ and $\tau_S = 0.50$. Non-dropped samples run one gradient-carrying forward at the sampled τ for the loss plus two no-gradient forwards at τ_S for the action effect, adding no backward pass and no external model; dropped samples train the null-action branch with uniform flow matching. We report each benchmark's official metrics and, where feasible, mean and standard deviation across repeated runs, wall-clock time, peak memory, and steps to convergence.

3.2 Main Comparison Across Action-Conditioned World Model

To comprehensively assess the generality of CAER as a training paradigm, we compare it with uniform MSE across four heterogeneous action-conditioned tasks and their corresponding benchmarks. Each pair uses the same initialization, data, action interface, and optimization schedule, with the objective as the only difference. Tables 2, 3, and 4 report the official metrics of iWorld-Bench, WorldArena, and VBench, covering visual and motion quality, physical consistency, controllability, and semantic

Factor	Value	WorldArena ↑	iWorldBench ↑
Action dropout	5%	62.83	0.5447
	10%	**63.13**	**0.6614**
	15%	61.91	0.5797
	20%	61.40	0.5470
Fixed time τ_S	0.25	61.09	0.5652
	0.50	**63.13**	**0.6614**
	0.75	62.27	0.6167

Table 5: Sensitivity of CAER to action dropout and fixed noise τ_S on RoboTwin (WorldArena) and camera control (iWorld-Bench). Bold entries denote the default configuration.

alignment.

Across all four AC-WM settings, CAER consistently outperforms the matched uniform-MSE baseline under the same backbone, data, and training schedule. Gains appear not only in aggregate scores but also across fine-grained dimensions that stress action following, interaction fidelity, and physical consistency, indicating that action-causal reweighting systematically redirects capacity toward tokens that determine controllable dynamics rather than background residuals. Because the improvement holds for robot manipulation (LIBERO and RoboTwin), camera control, and pose-conditioned generation alike, these results support CAER as a more general and better-suited training objective for action-conditioned world models than space–time-uniform MSE.

Figure 3 confirms that the loss concentrates on action-relevant regions across control modalities while downweighting already-modeled content. Benchmark-specific qualitative comparisons are provided in Appendix A: camera control is shown in Figure 5, RoboTwin in Figure 6, PoseAnything in Figure 7, and LIBERO in Figure 8. These examples provide a closer view of how CAER shifts supervision toward action-relevant regions under different control interfaces. Quantitatively, CAER improves the score from 0.6412 to 0.6614 on camera control, from 57.66 to 61.79 on LIBERO, from 62.35 to 63.13 on RoboTwin, and from 0.7422 to 0.7746 on PoseAnything. The largest gains occur in motion- and interaction-sensitive dimensions, including dynamic degree, flow quality, content consistency, and imaging quality. Several strong appearance or trajectory metrics remain unchanged or slightly decrease. This suggests that CAER reallocates capacity from easy, action-independent content toward unresolved action-conditioned dynamics rather than optimizing every token uniformly, improving world-model performance.

3.3 Training Hyperparameter Analysis

As shown in Table 5, we next analyze the sensitivity of CAER to action dropout and the fixed-noise operating point. For action dropout, we compare $p_{\text{drop}} \in \{5\%, 10\%, 15\%, 20\%\}$ while fixing $\tau_S = 0.50$. For the effect-map operating point, we compare $\tau_S \in \{0.25, 0.50, 0.75\}$ while fixing $p_{\text{drop}} = 10\%$. Changing one factor at a time separates null-branch quality from noise-level sensitivity.

The results show a clear trade-off in the action dropout rate. A very small dropout rate insufficiently trains the null-action branch, whereas an excessively large rate causes action-irrelevant regions to dominate the reweighting signal. Empirically, approximately 10% provides a reasonable balance.

The results also highlight the importance of selecting an appropriate noise level τ_S. Extremely small or large values either suppress the difference between the two branches or weaken scene-specific localization. In comparison, $\tau_S = 0.5$ preserves both action dependence and scene context, producing the most informative action-effect map.

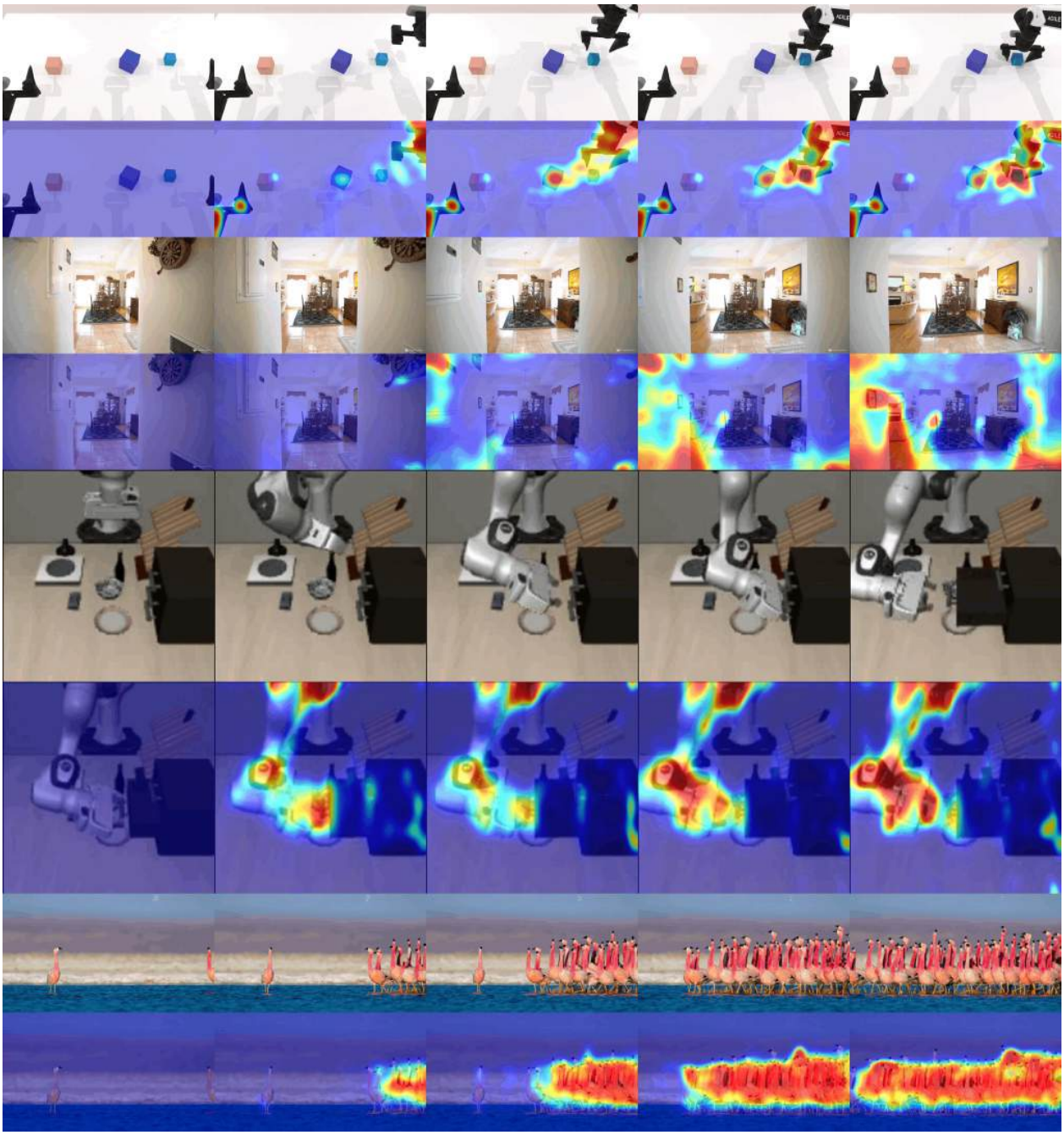

Figure 3: Qualitative Results. Generated videos and their CAER loss-supervision maps across four distinct action-conditioning modalities. Darker red regions indicate larger loss values, whereas darker purple regions indicate smaller loss values.

3.4 Toy Study: Score Evolution During Training

Finally, we evaluate intermediate checkpoints throughout training on both benchmarks to characterize how the two objectives trade off over optimization. Figure 4 plots iWorld-Bench on camera control and WorldArena on RoboTwin, with CAER and uniform MSE sharing the same initialization, data order, and schedule. The four curves exhibit an early-lead-to-late-lead crossover pattern. At the earliest checkpoints, uniform MSE often scores higher, since spreading the coefficient mass over all tokens quickly reduces abundant background residuals that appearance-driven metrics reward.

CAER instead allocates part of that mass to the harder interaction tokens. As the action-conditional dynamics become better learned, CAER catches up and generally overtakes uniform MSE on both benchmarks. Although the curves retain local fluctuations, CAER maintains a positive late-stage margin, while uniform MSE largely plateaus with interaction errors remaining unresolved. This crossover is consistent with the behavior predicted by the self-improving loop: a sharper θ yields a sharper effect map and directs more of the fixed coefficient mass toward what remains unresolved.

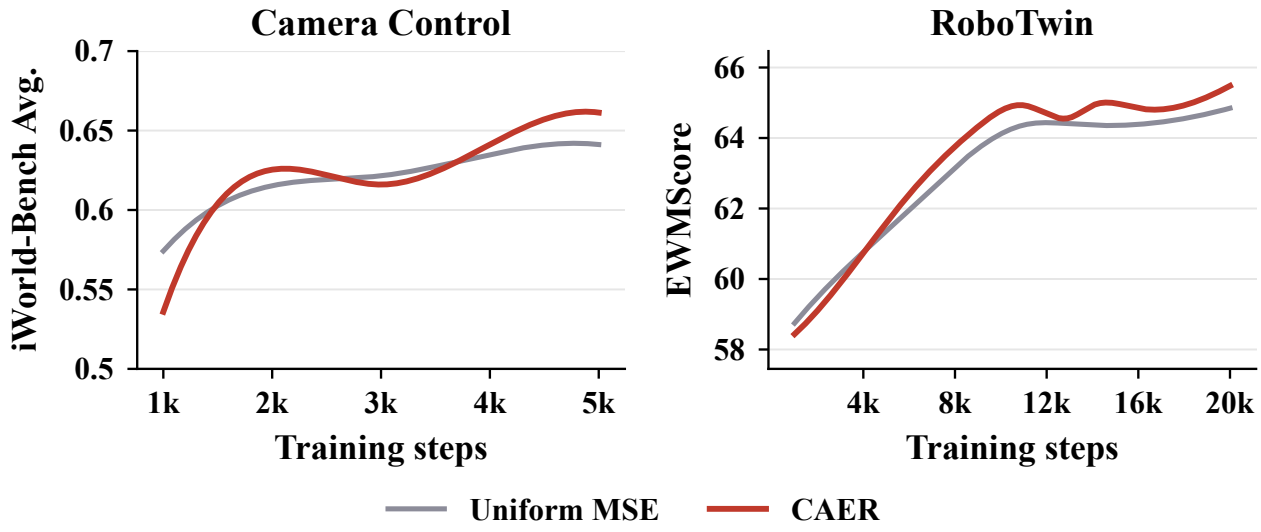

Figure 4: Score evolution across intermediate training checkpoints. iWorld-Bench on camera control and WorldArena on RoboTwin are shown under CAER and uniform MSE. Uniform MSE often leads at the earliest checkpoints, whereas CAER catches up and becomes higher later, while retaining small local fluctuations throughout training.

4 Related Work

Video-generative world models. Large video diffusion and flow-matching models, including Wan [1], HunyuanVideo [2], CogVideoX [21], and Cosmos [3], provide the visual and temporal priors on which many recent world models are built. Action-conditioned systems span camera trajectories, navigation commands, game controls, robot actions, and human motion. Genie [22], Matrix-Game [23–25], HY-World [26, 27], Cosmos-Predict [4, 28], and Worldscape-MoE [6] study interactive scene exploration, while DreamGen [29], RoboDreamer [30], Hand2World [31], Generated Reality [32], WorldVLN [7], and WorldScape Policy [8] extend video world models to robot, hand, body, and navigation control. Despite their different interfaces, these methods generally retain a space–time-uniform denoising objective, allowing sparse action consequences to be overwhelmed by background tokens. CAER addresses this shared optimization problem by redistributing supervision after action injection rather than introducing another control interface.

External spatial supervision for interaction-aware reweighting. External models such as SAM [10] and RAFT [11] provide segmentation and motion cues for localizing dynamic regions. MotiF [33] reweights video objectives with optical-flow heatmaps; BridgeV2W [34], Mask World Model [35], and Mask2Real-WM [36] use rendered or semantic masks; and ChronoDreamer [37] learns from simulator contact maps. Although effective, these pipelines require per-frame preprocessing, rendering, or simulation, and their visual proxies need not coincide with action causality. CAER instead extracts interaction-aware weights online from the model's own action response without external spatial annotations.

5 Conclusion and Future Work

We propose CAER, a general world-model training objective that identifies action-causal regions by contrasting predictions with and without actions, then emphasizes unresolved interactions while preserving the overall coefficient mass. Across four heterogeneous action-conditioned settings, CAER consistently outperforms uniform MSE and develops increasingly accurate interaction focus as training progresses. Building on this framework, future work will investigate broader mechanisms and effects of action causality in world models, including how interventions shape learned dynamics, generalization, and controllability, how such reweighting scales to longer horizons and larger backbones, and how this contrast can supply signals for evaluation and data selection.

References

[1] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[2] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[3] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

[4] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[5] GigaWorld Team. Gigaworld-0: An open-access world model for embodied ai. *arXiv preprint*, 2025.

[6] Jianjie Fang, Yongyan Xu, Ziyou Wang, Chen Gao, Yuchao Huang, Zhaolu Wang, Rongze Tang, Mingyuan Jia, Baining Zhao, Weichen Zhang, Xin Zhang, Haisheng Su, Yu Shang, Wei Wu, Xinlei Chen, and Yong Li. Worldscape-MoE: A unified mixture-of-experts world model for scalable heterogeneous action control. *arXiv preprint arXiv:2607.03964*, 2026.

[7] Baining Zhao, Jiacheng Xu, Weicheng Feng, Xin Zhang, Zhaolu Wang, Haoyang Wang, Shilong Ji, Ziyou Wang, Jianjie Fang, Zhiheng Zheng, Weichen Zhang, Yu Shang, Wei Wu, Chen Gao, Xinlei Chen, and Yong Li. WorldVLN: Autoregressive world action model for aerial vision-language navigation. *arXiv preprint arXiv:2605.15964*, 2026.

[8] Haisheng Su, Zongdai Liu, Xin Jin, Haoxuan Dou, Chengming Hu, Baorun Li, Zhanwang Liu, Ruiyan Xu, Jianjie Fang, Xin Zhang, Zhenjie Yang, Xue Yang, Chen Gao, Junchi Yan, Yong Li, and Wei Wu. WorldScape Policy 2.0: Empowering steerable world action modeling with reasoning-augmented memory. *arXiv preprint arXiv:2607.18840*, 2026.

[9] Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, and Yong Li. iworld-bench: A benchmark for interactive world models with a unified action generation framework. *arXiv preprint arXiv:2605.03941*, 2026.

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[11] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020.

[12] Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 2022.

[13] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[15] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

[16] Yu Shang et al. Worldarena: A benchmark for evaluating embodied world models. *arXiv preprint*, 2026.

[17] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

[18] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[19] Ruiyan Wang, Teng Hu, Kaihui Huang, Zihan Su, Ran Yi, and Lizhuang Ma. Poseanything: Universal pose-guided video generation with part-aware temporal coherence. *arXiv preprint arXiv:2512.13465*, 2025.

[20] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[21] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[22] Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

[23] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

[24] Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

[25] Zile Wang, Zexiang Liu, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

[26] Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

[27] HY-World Team et al. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026.

[28] NVIDIA. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

[29] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

[30] Gaoyue Zhou, Victoria Dean, Mohan Kumar Srirama, Aravind Rajeswaran, Jyothish Pari, Kyle Hatch, Aryan Jain, Tianhe Yu, Pieter Abbeel, and Lerrel Pinto. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

[31] Ziyou Wang et al. Hand2world: Hand-controllable world generation with egocentric videos. *arXiv preprint*, 2026.

[32] Tianyi Xie et al. Generated reality: Learning to generate interactive reality from human video. *arXiv preprint*, 2026.

[33] Shijie Wang, Samaneh Azadi, Rohit Girdhar, Saketh Rambhatla, Chen Sun, and Xi Yin. MotiF: Making text count in image animation with motion focal loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7773–7783, 2025.

[34] Yixiang Chen, Peiyan Li, Jiabing Yang, Keji He, Xiangnan Wu, Yuan Xu, Kai Wang, Jing Liu, Nianfeng Liu, Yan Huang, and Liang Wang. BridgeV2W: Bridging video generation models to embodied world models via embodiment masks. *arXiv preprint arXiv:2602.03793*, 2026.

[35] Yunfan Lou, Xiaowei Chi, Xiaojie Zhang, Zezhong Qian, Chengxuan Li, Rongyu Zhang, Yaoxu Lyu, Guoyu Song, Chuyao Fu, Haoxuan Xu, Pengwei Wang, and Shanghang Zhang. Mask world model: Predicting what matters for robust robot policy learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

[36] Riccardo Orion Feingold, Davide Liconti, Chenyu Yang, and Robert K. Katzschmann. Mask2Real-WM: Segmentation masks as a sim-to-real bridge for controllable dexterous world models. In *Conference on Robot Learning*, 2026.

[37] Zhenhao Zhou and Dan Negrut. ChronoDreamer: Action-conditioned world model as an online simulator for robotic planning. *arXiv preprint arXiv:2512.18619*, 2025.

A Additional Qualitative Comparisons

This appendix provides benchmark-specific qualitative comparisons between uniform MSE and CAER across the four action-conditioned settings. Each figure presents representative conditioning or original frames together with the corresponding supervision maps produced by uniform MSE and CAER. The same color convention as in Figure 3 is used throughout: warmer colors indicate larger loss responses, whereas cooler colors indicate smaller responses. Overall, CAER produces more spatially concentrated responses around action-relevant regions, while uniform MSE distributes supervision more broadly across the scene.

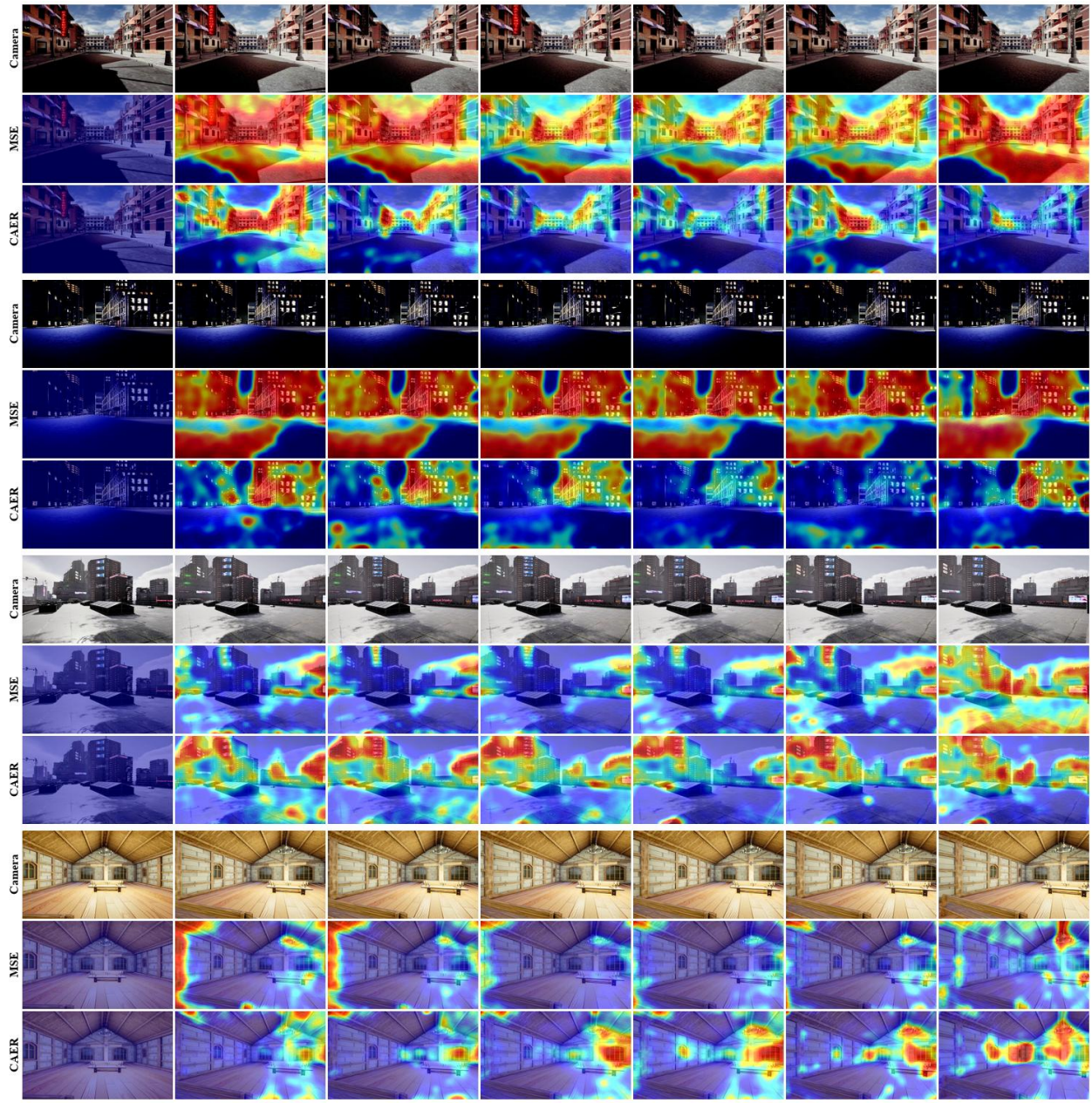

Figure 5: Camera-control qualitative comparison. Representative camera-control examples and their supervision maps under uniform MSE and CAER. Compared with the spatially broader responses of uniform MSE, CAER places more emphasis on regions associated with camera-induced motion and scene changes while suppressing already modeled background content.

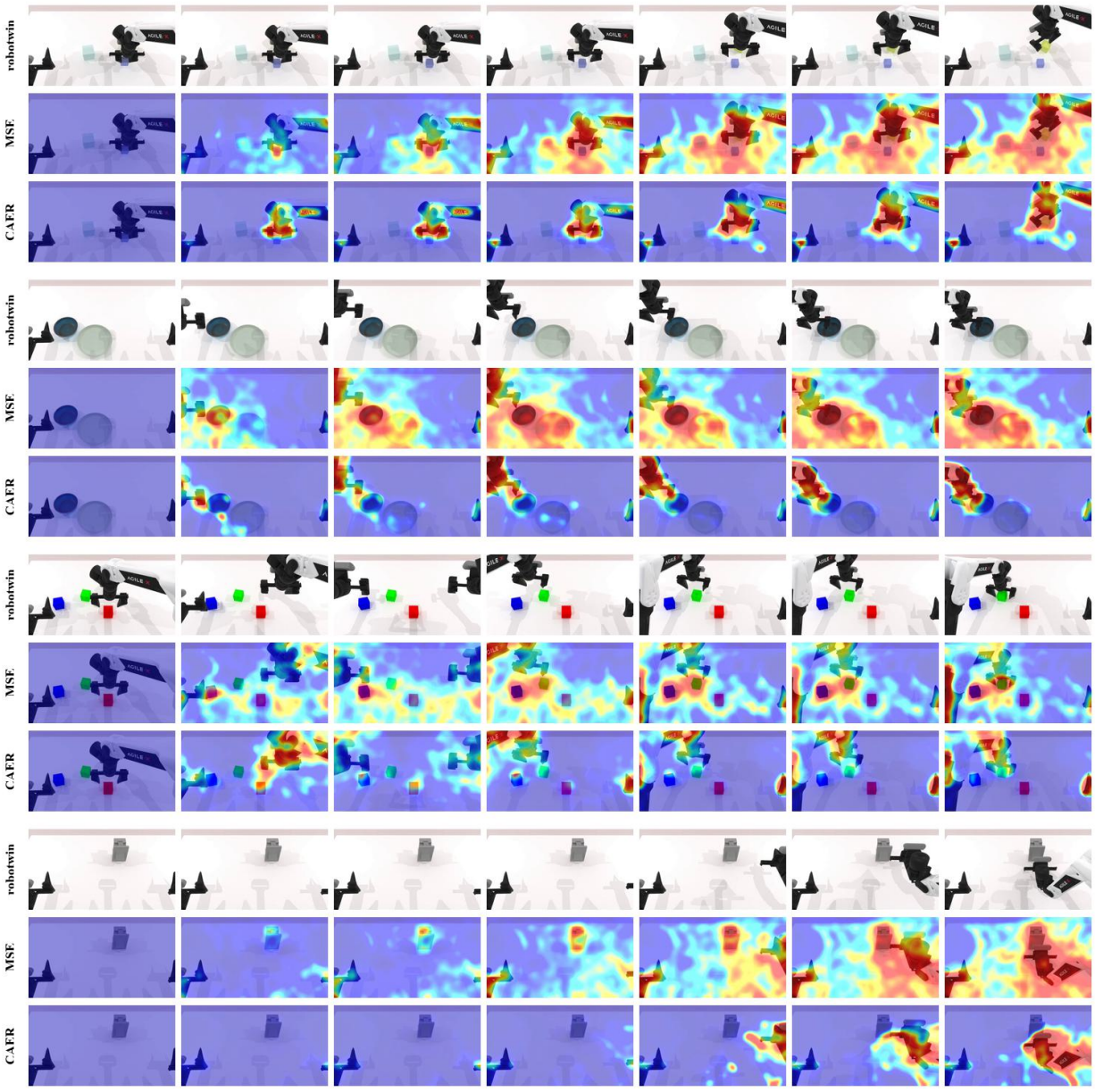

Figure 6: RoboTwin arm-manipulation qualitative comparison. Representative RoboTwin manipulation examples and the corresponding supervision maps under uniform MSE and CAER. CAER more consistently concentrates responses around the robot end-effector, manipulated objects, and interaction or contact regions, while uniform MSE assigns substantial responses to broader scene regions.

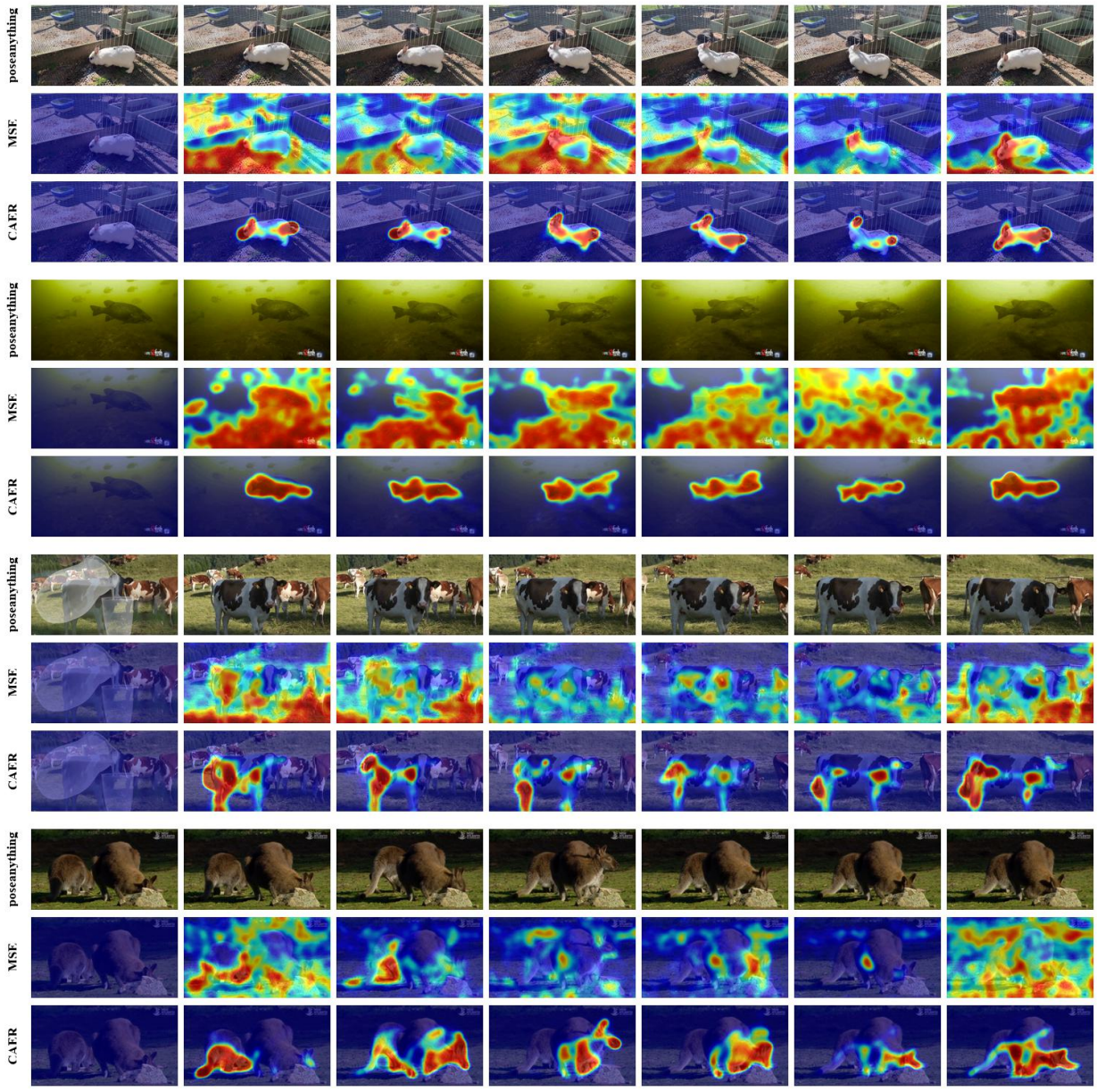

Figure 7: PoseAnything qualitative comparison. Representative pose-conditioned generation examples and their supervision maps under uniform MSE and CAER. The CAER maps are more focused on articulated subjects and pose-bearing regions, whereas uniform MSE exhibits more diffuse responses over the surrounding visual content.

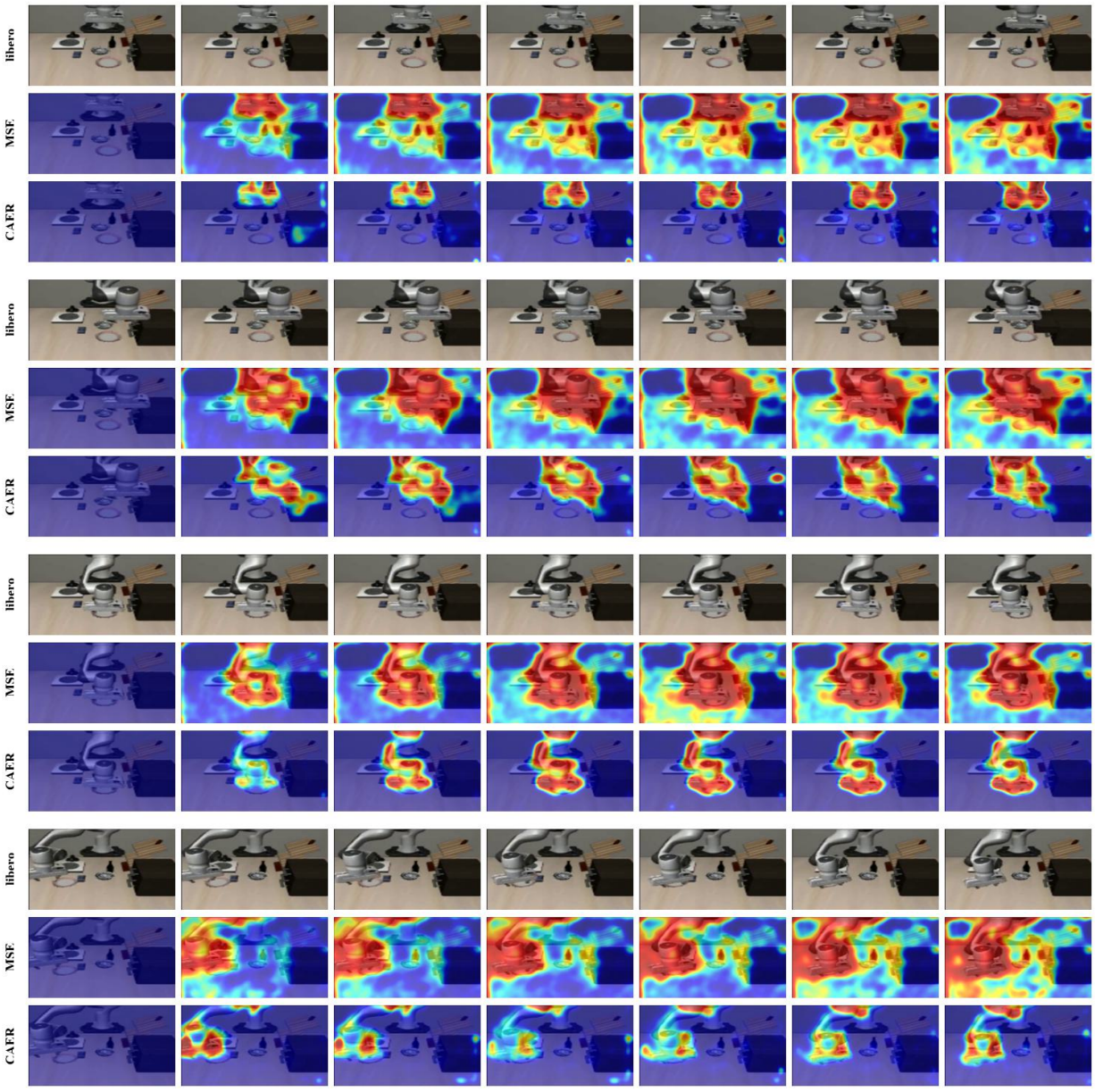

Figure 8: LIBERO manipulation qualitative comparison. Representative LIBERO robot-manipulation examples and the corresponding supervision maps. CAER emphasizes the robot, manipulated object, and relevant workspace involved in the action, while uniform MSE spreads supervision more broadly across the scene.