

# DIAMOND-LoL: Enforcing Lieb-Robinson Locality in Diffusion World Models for Long-Horizon Consistency

## Supplementary Material

### A. Implementation Details for LoL Loss Components

This appendix provides the implementation details for the two key data-driven components of the Lieb-Robinson Locality Loss ( $\ell_{\text{LoL}}$ ): the estimation of the single-step light-cone radius,  $r_t$ , and the generation of the static-source mask,  $M$ . Both components are derived exclusively from the training data, requiring no external annotations or prior knowledge.

#### A.1. Estimation of the Data-Driven Light-Cone Radius ( $r_t$ )

The light-cone radius,  $r_t$ , is intended to learn from data the maximum plausible displacement of a pixel boundary within a single environment step. An accurate  $r_t$  is crucial for correctly constraining the model’s dynamic predictions. We employ a statistical method based on the true state transitions in the training dataset to estimate this value. Specifically, for each ground-truth state transition sample  $(x_t, x_{t+1})$ , we compute its **Minimal Covering Radius**, defined as the smallest integer  $r$  that satisfies the inclusion condition  $r_{\text{sample}} = \min\{r \in \mathbb{N} \mid \text{supp}(E(x_{t+1})) \subseteq \text{Dilate}_r(\text{supp}(E(x_t)))\}$ . To obtain a more robust and representative global radius, we use a **Quantile Upper Bound** method. We compute the minimal covering radius for a large subset of the training data to form a distribution  $\{r_{\text{sample}}\}$  and take its 95th percentile as the global light-cone radius  $r_t$ . This approach effectively filters outliers while ensuring the radius covers the vast majority of physically plausible movements. A single global  $r_t$  value determined by this method is used for all experiments.

#### A.2. Generation of the Static-Source Mask ( $M$ )

In environments such as Atari, non-dynamic elements like scoreboards and HUDs undergo instantaneous changes that do not follow finite-speed propagation laws. To avoid incorrectly penalizing the model for these events, we introduce a purely data-driven **static-source mask**,  $M$ , to identify and exclude these “source term” regions. The mask is generated via a frequency statistics procedure. We iterate through consecutive frame pairs  $(x_{t-1}, x_t)$  in the dataset to identify transient events, where a boundary appears at a pixel location  $p$  ( $E(x_t)[p] \geq \epsilon$ ) that is not a boundary in the previous frame ( $E(x_{t-1})[p] < \epsilon$ ). These events are accumulated in a 2D frequency map,  $F$ , where each count  $F[p]$  represents the number of times a boundary has appeared *ex nihilo* at that location. Finally, after processing the dataset, we generate

the binary mask  $M$  by thresholding this frequency map. A high-percentile threshold,  $T_{\text{freq}}$ , is computed from the non-zero frequencies, and the final mask is defined as:

$$M[p] = \begin{cases} 0 & \text{if } F[p] > T_{\text{freq}}, \\ 1 & \text{otherwise.} \end{cases} \quad (16)$$

This process automatically identifies and masks out regions like scoreboards that update independently, ensuring that the locality constraint is applied only to genuinely dynamic objects.

### B. Proof of the Zero-Loss Characterization

This section provides a rigorous proof for the zero-loss characterization of the Lieb-Robinson Locality Loss ( $\ell_{\text{LoL}}$ ) as stated in Section 4.3.1 of the main paper.

We begin by formally defining the core operators. Let the image space be defined on a discrete pixel grid  $\Lambda \subset \mathbb{Z}^2$ . An image is a function  $x : \Lambda \rightarrow \mathbb{R}^C$ .

- The boundary operator  $E(x)$  is defined as the  $\ell_1$ -norm of the discrete gradient,  $E(x) = \|\nabla x\|_1$ . The support of the boundary,  $\text{supp}(E(x))$ , is the set of pixels with non-zero boundary values:  $\text{supp}(E(x)) = \{p \in \Lambda : E(x)[p] > 0\}$ .
- The morphological dilation operator,  $\text{Dilate}_r(S)$ , expands a set of pixels  $S \subseteq \Lambda$  by a radius  $r \in \mathbb{N}$  using an  $\ell_\infty$ -ball as the structuring element. It is equivalent to the Minkowski sum  $S \oplus B_r$ , where  $B_r$  is an  $\ell_\infty$ -ball of radius  $r$ . We can define it as:  $\text{Dilate}_r(S) = \{q \in \Lambda : \text{dist}_\infty(q, S) \leq r\}$ .
- The static-source mask  $M \in [0, 1]^\Lambda$  is a tensor that down-weights pixels corresponding to static elements like scoreboards or HUDs, ensuring they are not penalized by the locality loss.

Let the finite-propagation set  $\mathcal{S}_{r_t}(x_t)$  be defined as:

$$\mathcal{S}_{r_t}(x_t) = \left\{ z : \text{supp}(E(z)) \subseteq \text{Dilate}_{r_t}(\text{supp}(E(x_t))) \right. \\ \left. \wedge \text{supp}(E(x_t)) \subseteq \text{Dilate}_{r_t}(\text{supp}(E(z))) \right\}.$$

For any predicted frame  $\hat{x}_{t+1}$ , the Lieb-Robinson Locality Loss is zero if and only if  $\hat{x}_{t+1}$  belongs to this set:

$$\ell_{\text{LoL}}(\hat{x}_{t+1}; x_t) = 0 \iff \hat{x}_{t+1} \in \mathcal{S}_{r_t}(x_t).$$

This equivalence is understood to hold within the dynamic regions of the image, where the static mask  $M$  is positive.

*Proof.* The proof proceeds in two directions.

( $\Rightarrow$ ): Assume  $\ell_{\text{LoL}}(\hat{x}_{t+1}; x_t) = 0$ . Since  $\ell_{\text{LoL}} = \ell_{\text{emerge}} + \ell_{\text{vanish}}$  and both terms are non-negative, this implies  $\ell_{\text{emerge}} = 0$  and  $\ell_{\text{vanish}} = 0$ . From the definition of  $\ell_{\text{emerge}}$  in Equation (11),  $\ell_{\text{emerge}} = 0$  means that for every pixel  $p \in \Lambda$ , the product  $(E(\hat{x}_{t+1}) \odot M)[p] \cdot (1 - \text{Dilate}_{r_t}(E(x_t)))[p]$  is zero. This implies that for any pixel  $p$  where  $E(\hat{x}_{t+1})[p] > 0$  and  $M[p] > 0$ , it must be that  $\text{Dilate}_{r_t}(E(x_t))[p] = 1$ . This is equivalent to the set inclusion:

$$\text{supp}(E(\hat{x}_{t+1})) \cap \{p | M(p) > 0\} \subseteq \text{Dilate}_{r_t}(\text{supp}(E(x_t))).$$

Similarly, from Equation (12),  $\ell_{\text{vanish}} = 0$  implies the reverse inclusion:

$$\text{supp}(E(x_t)) \cap \{p | M(p) > 0\} \subseteq \text{Dilate}_{r_t}(\text{supp}(E(\hat{x}_{t+1}))).$$

Since the mask  $M$  only serves to disregard static regions and does not alter the inclusion relationship within the dynamic regions, these two conditions together mean that  $\hat{x}_{t+1} \in \mathcal{S}_{r_t}(x_t)$ .

( $\Leftarrow$ ): Assume  $\hat{x}_{t+1} \in \mathcal{S}_{r_t}(x_t)$ . This means both set inclusions hold. The first inclusion,  $\text{supp}(E(\hat{x}_{t+1})) \subseteq \text{Dilate}_{r_t}(\text{supp}(E(x_t)))$ , guarantees that for any pixel  $p$  where  $E(\hat{x}_{t+1})[p] > 0$ , the term  $(1 - \text{Dilate}_{r_t}(E(x_t)))[p]$  must be 0. Therefore, the element-wise product inside the norm for  $\ell_{\text{emerge}}$  is identically zero, and thus  $\ell_{\text{emerge}} = 0$ . By the same logic, the second inclusion guarantees that  $\ell_{\text{vanish}} = 0$ . The sum is therefore  $\ell_{\text{LoL}} = 0$ .  $\square$

## C. Proof of the Preference for Modal Selection

This section provides the proof for the claim in Section 4.3.2 that the  $\ell_{\text{LoL}}$  loss resolves the mode-averaging problem.

Let the true next state  $x_{t+1}$  be generated from a mixture of discrete modes  $\{T_{\delta_k}(x_t)\}_k$ , where  $T_{\delta_k}$  is a transformation (e.g., translation) by a displacement vector  $\delta_k$ . Let  $r_t$  be the data-driven light-cone radius for the current step, determined by the ground-truth transition. Assume there exists at least one mode  $j$  in the mixture that is boundary-crossing, meaning its boundary support lies partially outside the light-cone of the current state:

$$\text{supp}(E(T_{\delta_j}(x_t))) \not\subseteq \text{Dilate}_{r_t}(\text{supp}(E(x_t))).$$

For any convex combination of the modes (a pixel-space average)  $\bar{x} = \sum_k \pi_k T_{\delta_k}(x_t)$  with  $\pi_k > 0$  and  $\sum_k \pi_k = 1$ , we have  $\ell_{\text{LoL}}(\bar{x}; x_t) > 0$ . Conversely, any single reachable mode  $T_{\delta_k}(x_t)$  that is not boundary-crossing satisfies  $\ell_{\text{LoL}} = 0$ . Therefore, the minimization of the LoL loss favors selecting a single mode over the average:

$$\arg \min_y \ell_{\text{LoL}}(y; x_t) \subseteq \{T_{\delta_k}(x_t)\}_k.$$

*Proof.* By the linearity of the gradient operator, the boundary map of the averaged prediction is  $\nabla \bar{x} = \sum_k \pi_k \nabla T_{\delta_k}(x_t)$ . Assuming the boundary supports of the different modes are sufficiently separated on the discrete pixel grid, their pixel-wise sum will not result in significant cancellations (i.e., we ignore overlaps of measure zero). Thus, the support of the average’s boundary is approximately the union of the individual supports:

$$\text{supp}(E(\bar{x})) \approx \bigcup_k \text{supp}(E(T_{\delta_k}(x_t))).$$

By our proposition’s premise, there exists a boundary-crossing mode  $j$ . From the union property above, the support of the average,  $\text{supp}(E(\bar{x}))$ , must contain the support of this mode,  $\text{supp}(E(T_{\delta_j}(x_t)))$ . Because  $\text{supp}(E(T_{\delta_j}(x_t)))$  is not a subset of  $\text{Dilate}_{r_t}(\text{supp}(E(x_t)))$ , it follows that  $\text{supp}(E(\bar{x}))$  is also not a subset of  $\text{Dilate}_{r_t}(\text{supp}(E(x_t)))$ .

This means there exists a set of pixels with positive measure where  $E(\bar{x}) > 0$  but  $(1 - \text{Dilate}_{r_t}(E(x_t))) > 0$ . From the definition in Equation (11), this directly implies that  $\ell_{\text{emerge}}(\bar{x}; x_t) > 0$ , and consequently,  $\ell_{\text{LoL}}(\bar{x}; x_t) > 0$ .

For any single mode  $T_{\delta_k}(x_t)$  that is not boundary-crossing, both the emerge and vanish conditions for the finite-propagation set  $\mathcal{S}_{r_t}(x_t)$  can be satisfied, and by Proposition A, its  $\ell_{\text{LoL}}$  will be zero.  $\square$

The radius  $r_t$  is empirically estimated from the ground-truth transition  $(x_t, x_{t+1})$  to be the minimal covering radius. It therefore only characterizes the maximum displacement of the single mode that was actually realized. For any other unrealized mode in the mixture, if any part of its boundary moves further than this minimal radius, it will be considered superluminal or boundary-crossing and will trigger the  $\ell_{\text{emerge}}$  penalty. This is the precise mechanism by which LoL encourages the selection of a single, physically plausible mode over an average of multiple modes. This refined condition is more rigorous than simply stating that the distance between modes  $\|\delta_i - \delta_j\|_\infty > r_t$ .

## D. Proof of the Linear Bound on Long-Horizon Error

This section provides the proof sketch for the claim in Section 4.3.3 that the  $\ell_{\text{LoL}}$  loss leads to at most linear error accumulation.

If at each step  $s$  of a rollout, the loss is bounded,  $\ell_{\text{LoL}}(\hat{x}_{s+1}; x_s) \leq \varepsilon$  for some small  $\varepsilon > 0$ , then the boundary deviation, measured by the Hausdorff distance  $d_H$ , satisfies:

$$d_H(\text{supp}(E(\hat{x}_{t+\tau})), \text{supp}(E(x_{t+\tau}))) \leq C_0 + C_1 \tau \varepsilon,$$

where the constants  $C_0$  and  $C_1$  depend only on the geometry of the dilation.

*Proof.* The proof proceeds in three steps, formalizing the intuition that local consistency implies global linear stability.

From the definition of the  $\ell_1$ -norm based losses in Equations (11) and (12), the condition  $\ell_{\text{LoL}} \leq \varepsilon$  implies that the total measure (a weighted pixel count) of the two defect sets is bounded by  $\varepsilon$ . These sets are:

$$\begin{aligned} A_{s+1}^{\text{out}} &:= \text{supp}(E(\hat{x}_{s+1})) \setminus \text{Dilate}_{r_s}(\text{supp}(E(x_s))) \\ A_s^{\text{miss}} &:= \text{supp}(E(x_s)) \setminus \text{Dilate}_{r_s}(\text{supp}(E(\hat{x}_{s+1}))) \end{aligned}$$

Intuitively, this means that, except for a small number of leaked or missing pixels, the boundary supports of  $\hat{x}_{s+1}$  and  $x_s$  are mutually covered by each other under dilation by  $r_s$ .

Let  $\hat{A}_s = \text{supp}(E(\hat{x}_s))$  and  $A_s = \text{supp}(E(x_s))$ . Using standard properties of morphological operations and the Hausdorff distance, one can establish a single-step recursive bound. In the absence of defects (i.e., if  $A_{s+1}^{\text{out}}$  and  $A_s^{\text{miss}}$  are empty), the error growth is bounded by the allowed radius  $r_s$ . The presence of defects can add an additional term, leading to an inequality of the form:

$$\begin{aligned} d_H(\hat{A}_{s+1}, A_{s+1}) &\leq d_H(\hat{A}_s, A_s) + r_s \\ &\quad + \kappa \cdot (\mathbf{1}[A_{s+1}^{\text{out}} \neq \emptyset] + \mathbf{1}[A_s^{\text{miss}} \neq \emptyset]), \end{aligned}$$

where  $\kappa > 0$  is a constant related to the diameter of the structuring element.

To move from the indicator function to a term linear in  $\varepsilon$ , we introduce a mild regularity assumption common for pixel-based environments: any contiguous boundary component must have a minimum area/perimeter, bounded below by some  $\eta > 0$ . This prevents infinitely small, zero-measure defect clusters. Under this assumption, if the total measure of defects is bounded by  $\varepsilon$ , the total number of distinct defect clusters is bounded by  $\varepsilon/\eta$ . Summing the single-step inequality from step 2 over a horizon of  $\tau$  steps, the contribution from the defect indicators becomes proportional to  $\tau \cdot (\varepsilon/\eta)$ . This yields the final recursive bound:

$$d_H(\hat{A}_{t+\tau}, A_{t+\tau}) \leq d_H(\hat{A}_t, A_t) + \sum_{s=t}^{t+\tau-1} r_s + \frac{\kappa}{\eta} \tau \varepsilon.$$

By defining  $C_0 = d_H(\hat{A}_t, A_t) + \sum_{s=t}^{t+\tau-1} r_s$  and  $C_1 = \kappa/\eta$ , we arrive at the claimed linear bound.  $\square$