Rethinking Visual Rearrangement from A Diffusion Perspective

Supplementary Material

1. Theoretical Analysis

Since visual rearrangement tasks involve randomly shuffling the scene configuration, we assume that the state changes in the scene can be viewed as the effect of random external forces acting under physical constraints, therefore, the changing process obeys the Langevin dynamics [6], which can be formulated as:

$$\mathbf{X}_t = \mathbf{X}_{t-1} + \frac{\epsilon}{2}\nabla_{\mathbf{x}} \log p(\mathbf{X}_{t-1}) + \sqrt{\epsilon}\mathbf{z}_t \tag{1}$$

where $\mathbf{X}_t$ represents the state of scene at time step t, $\epsilon > 0$ is a fixed step size, $\nabla_{\mathbf{x}} \log p(\mathbf{X}_{t-1})$ is the score function[6] of the probability density $p(\mathbf{X}_{t-1})$ and $\mathbf{z}_t \sim \mathcal{N}(0, I)$. Since the visual rearrangement task is defined as a Markov random process[1], there is a general description in the form of Fokker-Planck equation. As Langevin dynamics describes discrete changes in the distribution of the scene, we can derive the corresponding Fokker–Planck equation[4], which describes continuous changes in the distribution of the scene. From the Fokker–Planck equation with the drift term equal to zero, we can get the diffusion equation in the form of Fick's second law:

$$\frac{\partial P(\mathbf{X}_t)}{\partial t} = D\nabla^2 P(\mathbf{X}_t) \tag{2}$$

where D denotes the diffusion coefficient, ∇^2 denotes the Laplacian operator indicating the changes of probability distribution over space.

We now proceed to deduce that the change in information entropy also conforms to the diffusion equation. The Shannon information entropy of the scene can be formulated as:

$$H(\mathbf{X}_t) = -\sum^{\mathbf{X}} P(\mathbf{X}_t) \log P(\mathbf{X}_t) \tag{3}$$

We first calculate the rate at which the information entropy changes over time. We take the time derivative of information entropy:

$$\begin{aligned}
\frac{dH}{dt} &= -\frac{d}{dt}\sum^{\mathbf{X}} P(\mathbf{X}_t) \log P(\mathbf{X}_t) \\
&= -\sum^{\mathbf{X}} \left(\frac{\partial P(\mathbf{X}_t)}{\partial t} \log P(\mathbf{X}_t) + P(\mathbf{X}_t) \frac{\partial \log P(\mathbf{X}_t)}{\partial t} \right) \\
&= -\sum^{\mathbf{X}} \left(\frac{\partial P(\mathbf{X}_t)}{\partial t} \log P(\mathbf{X}_t) + \frac{P(\mathbf{X}_t)}{P(\mathbf{X}_t)} \frac{\partial P(\mathbf{X}_t)}{\partial t} \right) \\
&= -\sum^{\mathbf{X}} D\nabla^2 P(\mathbf{X}_t) \log P(\mathbf{X}_t) - \sum^{\mathbf{X}} \frac{\partial P(\mathbf{X}_t)}{\partial t} \\
&= -D\sum^{\mathbf{X}} \nabla^2 P(\mathbf{X}_t) \log P(\mathbf{X}_t)
\end{aligned} \tag{4}$$

Then we calculate the second derivative of information entropy over space:

$$\begin{aligned}
\nabla^2 H(\mathbf{X}_t) &= -\sum^{\mathbf{X}} \nabla^2 \left(P(\mathbf{X}_t) \log P(\mathbf{X}_t) \right) \\
&= -\sum^{\mathbf{X}} \nabla \left(\frac{\partial P(\mathbf{X}_t)}{\partial x} \log P(\mathbf{X}_t) + P(\mathbf{X}_t) \frac{\partial \log P(\mathbf{X}_t)}{\partial x} \right) \\
&= -\sum^{\mathbf{X}} \left(\nabla^2 P(\mathbf{X}_t) \log P(\mathbf{X}_t) + \frac{\partial P(\mathbf{X}_t)}{\partial x} \frac{\partial \log P(\mathbf{X}_t)}{\partial x} \right)
\end{aligned} \tag{5}$$

Since the system is in a steady state or near equilibrium during the diffusion process, the speed of change of the system is very slow compared to the steady state, the derivative of $P(\mathbf{X}_t)$ over space can be approximated as the original function:

$$\begin{aligned}
\frac{\partial P(\mathbf{X}_t)}{\partial x} \frac{\partial \log P(\mathbf{X}_t)}{\partial x} &= \frac{\partial P(\mathbf{X}_t)}{\partial x} \frac{\partial \log \nabla P(\mathbf{X}_t)}{\partial x} \\
&= \frac{\nabla P(\mathbf{X}_t)}{\nabla P(\mathbf{X}_t)} \nabla^2 P(\mathbf{X}_t) \\
&= \frac{1}{D} \frac{\partial P(\mathbf{X}_t)}{\partial t}
\end{aligned} \tag{6}$$

By utilizing Eq. (6), we can get:

$$\nabla^2 H(\mathbf{X}_t) = -\sum^{\mathbf{X}} \nabla^2 P(\mathbf{X}_t) \log P(\mathbf{X}_t) \tag{7}$$

which results in:

$$\frac{\partial H(\mathbf{X}_t)}{\partial t} = D\nabla^2 H(\mathbf{X}_t) \tag{8}$$

In nonequilibrium thermodynamics, the diffusion process is an entropy-driven relaxation process. Correspondingly, the disruption and restoration of objects in the room reflect the changes of the scene in information entropy, which is likewise an entropy-driven process and the process is proven to conform to the diffusion equation. Hence, the rearrangement task can be modeled by the diffusion process.

2. Computational Overhead

We conduct experiments on test set of our self-built dataset and report the average time taken by the agent to complete each task. The experiments are conducted on 1 AMD EPYC 9654 CPU and 2 NVIDIA A40 GPU. Compared to TIDEE[5] and MaSS[7], the time consumption of CAVR[3] and our method is on the same order of magnitude and within an acceptable range. Our method demonstrates optimal performance metrics, achieving a 24.7% enhancement in Fixed Strict metric compared to the Kuhn-Munkres[2] matching approach, while requiring only a marginal additional computational time of 0.1 minutes per task.

Table 1. Computational overhead on test set of our dataset.

	Avg Time Per Task (minutes)	Fixed Strict (%) ↑
TIDEE[5]	15.6	10.8
MaSS[7]	18.3	9.7
CAVR[3]	**0.3**	17.0
Ours	0.6	**24.7**
Ours + *Direct*	0.5	10.0
Ours + *Kuhn-Munkres*	0.5	19.8

3. Limitation Analysis

We discuss the limitation of our method for the visual rearrangement task. Since we utilize Gaussian mixture model distributions to represent scene changes and predict transformations through diffusion-based approaches on distribution parameters, our focus is on changes in the positions and appearances of objects, which primarily correspond to pickable objects in the visual rearrangement task. In addition to pickable objects, openable objects also constitute a category of rearrangement targets. Openable objects remain fixed in position but can be opened to a certain extent. While our method effectively captures positional changes and successfully restores pickable objects, it lacks the capability to distinguish openable objects from other objects. While we adopt a solution of applying the `open` action to objects with minimal movement below a certain threshold, a more comprehensive discussion and analysis of handling openable objects remain an important direction for future work.

References

[1] Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, et al. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975*, 2020. 1

[2] Harold W Kuhn. The hungarian method for the assignment. *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*, page 29, 2009. 2

[3] Yuyi Liu, Xinhang Song, Weijie Li, Xiaohan Wang, and Shuqiang Jiang. A category agnostic model for visual rearrangment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16457–16466, 2024. 2

[4] Hannes Risken and Hannes Risken. *Fokker-planck equation*. Springer, 1996. 1

[5] Gabriel Sarch, Zhaoyuan Fang, Adam W Harley, Paul Schydlo, Michael J Tarr, Saurabh Gupta, and Katerina Fragkiadaki. Tidee: Tidying up novel rooms using visuo-semantic commonsense priors. In *European conference on computer vision*, pages 480–496. Springer, 2022. 2

[6] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1

[7] Brandon Trabucco, Gunnar Sigurdsson, Robinson Piramuthu, Gaurav S Sukhatme, and Ruslan Salakhutdinov. A simple approach for visual rearrangement: 3d mapping and semantic search. *arXiv preprint arXiv:2206.13396*, 2022. 2