A Geometric Perspective on Optimizing Vector Quantized Latent Diffusion Model for Image Restoration

Chen Hang[1], Haoming Chen[1], Xuwei Fang[2], Weisheng Xie[2], Xiangxiang Gao[2], Faming Fang[1], Guixu Zhang[1]*, Haichuan Song[1]

[1]East China Normal University, Shanghai, 200062 China
[2]Bestpay AI Lab, Shanghai, 200080 China
52265901002@stu.ecnu.edu.cn, 52265901012@stu.ecnu.edu.cn, fangxuwei@bestpay.com.cn, xieweisheng@bestpay.com.cn, gaoxiangxiang@bestpay.com.cn, fmfang@cs.ecnu.edu.cn, gxzhang@cs.ecnu.edu.cn, hcsong@cs.ecnu.edu.cn

Abstract

In this paper, we investigate the limitations of the Vector Quantized Latent Diffusion Model (VQ-LDM) in restoration tasks. We identify a performance gap between the Vector Quantization (VQ) and Diffusion Model components, manifested as a significant discrepancy between the reconstruction quality of ground truth images processed via VQ autoregression and degraded images restored by VQ-LDM. Through experiments, we attribute this gap primarily to the lack of robustness in the mapped points of VQ within the original VQ-LDM framework. To address this issue, we propose a geometric based optimization approach. First, we introduce a simple yet effective method, termed interpolation-based latent initial state optimization, which mitigates the performance gap by replacing the original mapped points with interpolated values, supported by theoretical analysis. Here, the latent initial state refers specifically to the input of the diffusion model. Building upon this, we further propose a Chebyshev center-based latent initial state optimization, an elegant theoretical solution from a geometric perspective, that further enhances restoration performance. Our improvements consistently achieve superior results across nine benchmark datasets.

Introduction

In recent years, Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Song, Meng, and Ermon 2020) have garnered significant attention and discussion in the field of image and video generation due to their high-quality generation capabilities. Concurrently, researchers have become curious about the performance of diffusion models in the domain of image and video restoration, leading to a substantial body of research in this area, particularly concerning the intersection of image restoration and diffusion models (Özdenizci and Legenstein 2023; Kawar et al. 2022; Luo et al. 2023; Yinhuai, Jiwen, and Jian 2022; Saharia et al. 2022). Standard diffusion models are primarily trained and inferred in the image space, which incurs significant resource and time costs. To reduce the time overhead, LDMs (Rombach et al. 2022) are proposed. Since LDMs are trained and inferred in the latent space, they require fewer resources compared to standard diffusion models for the same number of function evaluations (NFEs). However, in the field of image and video restoration, the research (Lin et al. 2023; Yang et al. 2023; Sun et al. 2024; Wang et al. 2023; Yue, Wang, and Loy 2023) on LDMs has primarily focused on incorporating additional modules or auxiliary networks to enhance restoration quality, with little attention given to investigating the fundamental causes of suboptimal restoration quality caused by LDMs. Through theoretical analysis and experimental results, We have found that the original VQ-LDM architecture is not the optimal architecture for restoration tasks. In vector quantization, the codebook geometrically constructs many Voronoi subregions (Voronoi cells). Any point within a Voronoi subregion has a distance to its corresponding Voronoi centroid (Voronoi site) that is less than the distance to the centroids of other Voronoi subregions, typically measured by the Euclidean distance. Using the values before and after vector quantization as input to the diffusion model, the results generated by the diffusion model may lead to remapping into other Voronoi subregions due to numerical errors. To reduce this error, we modify the corresponding mapping points after vector quantization.

First, we propose interpolation-based latent initial state optimization, a method that replaces the original mapped points with linearly interpolated values and uses them as inputs to the diffusion model. Despite being a minimal modification (requiring only minimal code modifications), this approach significantly improves restoration performance, supported by our theoretical analysis. Notably, it achieves train-free performance gains when applied to existing VQ-LDM method. Building on this discovery, we further develop Chebyshev center-based latent initial state optimization using linear programming from convex optimization. Experiments demonstrate that this method exhibits superior robustness and represents an elegant theoretical solution from a geometric perspective. Our main contributions are summarized as follows:

- We propose interpolation-based latent initial state optimization, a simple yet effective modification that significantly enhances the restoration performance of VQ-LDM, accompanied by theoretical analysis.
- Leveraging linear programming, we further introduce Chebyshev center-based latent initial state optimization,

*Corresponding Author

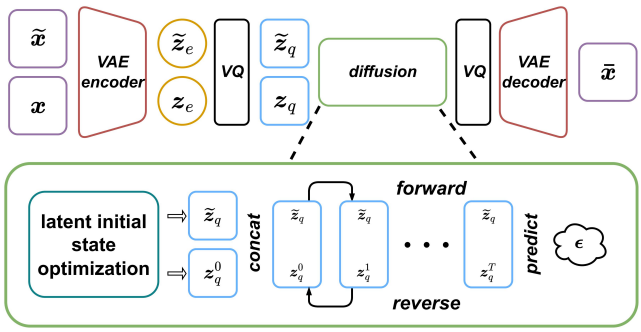

Figure 1: The workflow of VQ-LDM for image restoration.

which demonstrates superior robustness and achieves even better restoration performance compared to the interpolation-based approach.

- We conduct extensive experiments on image restoration tasks, validating our methods across nine datasets encompassing both images and videos.

Related Work

LDM has been widely applied to restoration tasks. DiffBIR (Lin et al. 2023) enhances restoration performance by using a preprocessing module and a conditional control module. Pixel-aware Stable Diffusion (Yang et al. 2023) introduces Pixel-Aware Cross Attention and incorporates additional high-level visual supervision to improve restoration performance. StableSR (Wang et al. 2023) introduces a trainable controllable feature wrapping module to navigate the trade-off between fidelity and realism. CoSeR (Sun et al. 2024) introduces priors from low-resolution images, CLIP's semantic priors, and reference images generated by the diffusion model to assist in image restoration. ResShift (Yue, Wang, and Loy 2023) accelerates image restoration by adjusting the white noise in the final diffusion step for low-quality images. As a VQ-LDM-based method, ResShift also serves as a testbed for our interpolation-based latent initial state optimization in later sections. Remarkably, this integration requires only minimal code modifications to the original ResShift implementation.

Motivation

First, we present the operational pipeline of the VQ-LDM framework in restoration tasks, with the detailed workflow depicted in Figure 1. Where $\bar{\boldsymbol{x}}$ represent the recovered image, $\widetilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ represent the degraded image and ground truth. $\boldsymbol{z}_e/\widetilde{\boldsymbol{z}}_e$ and $\boldsymbol{z}_q/\widetilde{\boldsymbol{z}}_q$ represent the pre-quantization and post-quantization tensors in the vector quantization process. $\boldsymbol{z}_q^T$ follow a Gaussian distribution. ϵ represent the predicted noise. In the original VQ-LDM, the typical practice is to directly utilize the vector quantized results $\boldsymbol{z}_q/\widetilde{\boldsymbol{z}}_q$ as the latent initial state of the diffusion model. Our proposed method primarily incorporates an latent initial state optimization module prior to the diffusion model, where the modified $\boldsymbol{z}_q/\widetilde{\boldsymbol{z}}_q$ are fed into the diffusion process to mitigate the performance gap in VQ-LDM. The latent initial state refers specifically to the input of the diffusion model. The interpolation-based latent initial state optimization replaces the original $\boldsymbol{z}_q/\widetilde{\boldsymbol{z}}_q$ with a linear interpolation between $\boldsymbol{z}_e/\widetilde{\boldsymbol{z}}_e$ and $\boldsymbol{z}_q/\widetilde{\boldsymbol{z}}_q$, formulated as: $\boldsymbol{z}_{lerp} = \lambda \boldsymbol{z}_e + (1-\lambda)\boldsymbol{z}_q$ where λ is the interpolation weight. The Chebyshev center-based latent initial state optimization will be formally introduced in subsequent sections. The additional vector quantization applied to the diffusion model's output in Figure 1 serves to correct the generated mapping points by aligning them with the codebook, which benefits the restoration task. This can be intuitively understood from two perspectives: (1) For pixels unaffected by degradation, they inherently require an auto-regressive solution, thus rigorous codebook alignment improves restoration quality; (2) For degraded pixels, maintaining rigorous codebook alignment also proves advantageous for training stability. During training (forward process), the diffusion model takes inputs $\boldsymbol{z}_q$ and $\widetilde{\boldsymbol{z}}_q$, while during inference (reverse process), it processes $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\widetilde{\boldsymbol{z}}_q$.

In the Introduction section, we characterize the codebook in VQ-LDM as a Voronoi diagram comprising multiple Voronoi subregions. The original VQ-LDM typically uses Voronoi centroids (i.e., the codebook's mapping points) as the latent initial state for the diffusion model. However, during inference, numerical errors inevitably cause deviations from these target mapping points, resulting in falls into neighboring Voronoi subregions. For a more intuitive understanding of this issue, we have included a conceptual sketch in the Figure 2. As evident from the Figure 2, the interpolation-based latent initial state optimization effectively prevents reconstruction mapping points from falling into neighboring Voronoi subregions, thereby enhancing both robustness and restoration performance. This method requires only minimal code modifications, representing a simple yet effective solution. The Chebyshev center-based latent initial state optimization further modifies $\boldsymbol{z}_q/\widetilde{\boldsymbol{z}}_q$ to the Chebyshev center of the Voronoi subregion. By definition, the Chebyshev center is the point maximally distant from all boundaries, equivalent to the center of the largest inscribed circle in two-dimensional polygons. While achieving enhanced robustness and restoration performance over the interpolation-based method, this approach introduces greater implementation complexity.

Latent Initial State Optimization

Figure 2 in the Motivation section visually illustrates the conceptual foundation of our latent initial state optimization. We now present rigorous theoretical analysis to support this approach, beginning with the interpolation-based method. As established in the Introduction section, the codebook in VQ-LDM forms a Voronoi diagram comprising multiple Voronoi subregions. Theorem 1 guarantees that each Voronoi subregion forms a convex hull, and moreover, any linear interpolation between two points within a Voronoi subregion remains confined to that subregion, ensuring $\boldsymbol{z}_{lerp}$ never deviates into neighboring Voronoi subregion. Proposition 1 further demonstrates that under numerical errors, adaptively adjusting the fixed latent initial state of original VQ-LDM using $\boldsymbol{z}_{lerp}$ prevents the diffusion model's outputs from crossing Voronoi boundaries. This significantly

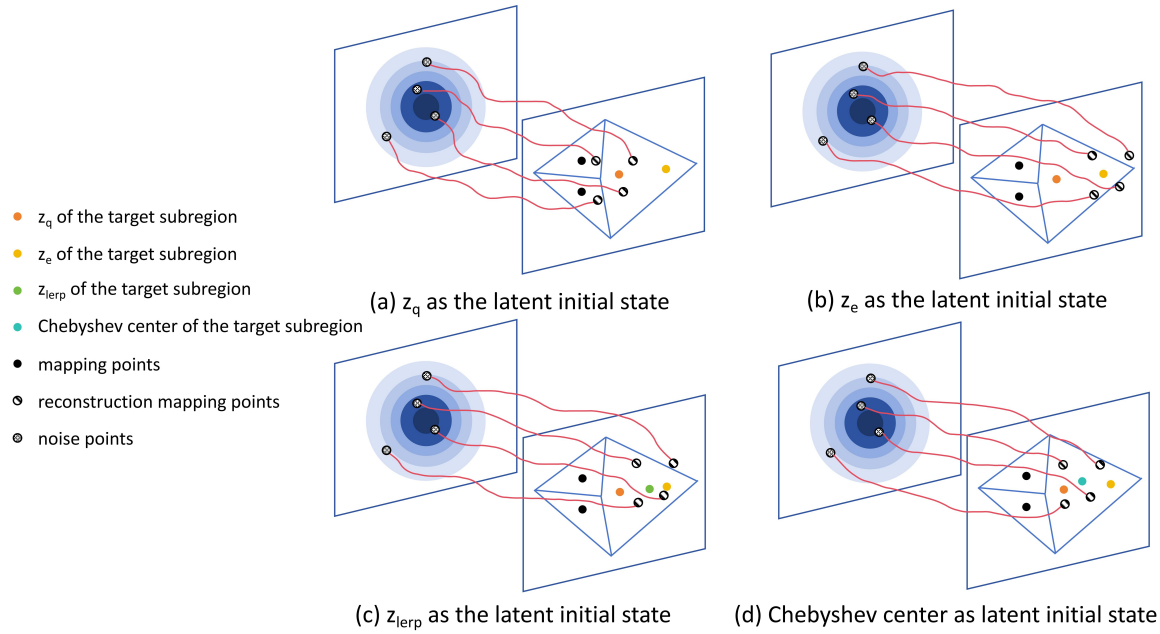

Figure 2: The definition of the latent initial state is the value after the VAE encoder, before the diffusion model. (a) to (d) represent the error-containing results obtained using different latent initial states. $\boldsymbol{z}_{lerp}$ represents the interpolation between $\boldsymbol{z}_e$ and $\boldsymbol{z}_q$. (c) and (d) correspond our methods.

enhances the robustness of subsequent vector quantization correction steps.

Theorem 1. *If there exists an n-dimensional Voronoi region V, then any subregion within V is an n-dimensional convex hull. We have that the interpolation $\boldsymbol{z}_{lerp}$ between $\boldsymbol{z}_e$ and $\boldsymbol{z}_q$ within any subregion remains within that subregion.*

Proof. Let us assume the existence of a subregion S within the Voronoi region V that does not satisfy the property of being a convex hull.

Let $f(p_1) = z_1$, $f(p_2) = z_1$, f denote the mapping function that projects points to their Voronoi subregion centroids. Here, p and z represent n-dimensional space points, where p is an arbitrary point within a Voronoi subregion, and z corresponds to the subregion's centroid.

Then $\exists p_3 = \lambda p_1 + (1-\lambda)p_2$, for $\lambda \in [0,1]$, such that $f(p_3) = z_2$.

Let $\overrightarrow{p_1 z_1} = \vec{a}$, $\overrightarrow{z_1 p_2} = \vec{b}$, $\overrightarrow{p_1 z_2} = \vec{c}$, $\overrightarrow{z_2 p_2} = \vec{d}$, $\overrightarrow{p_1 p_2} = \vec{e}$, $\overrightarrow{p_1 p_3} = \lambda\vec{e}$.

Since $|\vec{e}|^2 = (\vec{a}+\vec{b})\cdot(\vec{a}+\vec{b}) = |\vec{a}|^2 + |\vec{b}|^2 + 2\vec{a}\cdot\vec{b}$, $|\vec{a}| < |\vec{c}|$, $|\vec{b}| < |\vec{d}|$. Where $\cdot$ denotes the inner product.

Now,

$$
\begin{aligned}
|\overrightarrow{z_1 p_3}|^2 =& |\lambda\vec{e} - \vec{a}|^2 = |\lambda(\vec{a}+\vec{b}) - \vec{a}|^2 = |(\lambda-1)\vec{a} + \lambda\vec{b}|^2 \\
=& |(\lambda-1)\vec{a}|^2 + |\lambda\vec{b}|^2 + 2\lambda(\lambda-1)\vec{a}\cdot\vec{b} \\
=& (1-\lambda)^2|\vec{a}|^2 + \lambda^2|\vec{b}|^2 - \lambda(1-\lambda)(|\vec{e}|^2 - |\vec{a}|^2 \\
& - |\vec{b}|^2) \\
=& (1-\lambda)|\vec{a}|^2 + \lambda|\vec{b}|^2 - \lambda(1-\lambda)|\vec{e}|^2.
\end{aligned}
$$

Similarly $|\overrightarrow{z_2 p_3}|^2 = (1-\lambda)|\vec{c}|^2 + \lambda|\vec{d}|^2 - \lambda(1-\lambda)|\vec{e}|^2$.

Then $|\overrightarrow{z_2 p_3}|^2 - |\overrightarrow{z_1 p_3}|^2 = (1-\lambda)(|\vec{c}|^2 - |\vec{a}|^2) + \lambda(|\vec{d}|^2 - |\vec{b}|^2) > 0$ holds true.

So, the original assumption does not hold, indicating that the Voronoi region V is always a convex hull. □

Due to the minimal modifications required by interpolation-based latent initial state optimization and for further investigation, we directly applied this method to ResShift (Yue, Wang, and Loy 2023). We conducted train-free validation on ResShift, which may not fully exploit the potential of interpolation-based latent initial state optimization, yet the results still align with Proposition 1 and show improved performance for ResShift, as presented in Table 1. It's evident that interpolation-based latent initial state optimization provides consistent improvements across all ResShift experiments, though the gains are modest due to the train-free validation. In subsequent experiments, we incorporated $\boldsymbol{z}_{lerp}$ into training, achieving more significant improvements. These results strengthened our confidence in latent initial state optimization and motivated our exploration for geometrically optimal solutions.

Chebyshev center-based latent initial state optimization emerged as our elegant theoretical solution through further research. Table 1 reveals the challenge in determining the optimal value for hyperparameter λ, as it varies across different Voronoi subregions and $\boldsymbol{z}_e$ conditions. To address this, we introduced linear programming from convex optimization to compute Chebyshev centers within Voronoi subregions, then replaced $\boldsymbol{z}_q$ with these centers as the latent initial state for the diffusion model, thereby achieving more ro-

bust results. The pseudocode implementation of Chebyshev center-based latent initial state optimization is presented in Algorithm 1.

Proposition 1. *In an n-dimensional convex hull S, let $\boldsymbol{z}_e$ and $\boldsymbol{z}_q$ be two coordinate points. If there exists an error radius R, then the interpolation point $\boldsymbol{z}_{lerp} = \lambda \boldsymbol{z}_e + (1-\lambda)\boldsymbol{z}_q$, where λ is a scalar between 0 and 1, has a greater probability of remaining within the convex hull S. This indicates a stronger robustness against errors in terms of the interpolation point's adherence to the convex hull.*

Proof. $\exists \boldsymbol{r_e}$ is a vector s.t.$|\boldsymbol{r_e}|$ is minimal, let $\boldsymbol{z}_e + \boldsymbol{r_e} + \boldsymbol{\varepsilon_e} \notin S$, where $\boldsymbol{r_e}, \boldsymbol{\varepsilon_e}$ are vectors and $|\boldsymbol{\varepsilon_e}| > 0$, $\boldsymbol{\varepsilon_e}$ is collinear with $\boldsymbol{r_e}$ and shares the same direction. $\exists \boldsymbol{r_q}$ is a vector s.t.$|\boldsymbol{r_q}|$ is minimal, let $\boldsymbol{z}_q + \boldsymbol{r_q} + \boldsymbol{\varepsilon_q} \notin S$, where $\boldsymbol{r_q}, \boldsymbol{\varepsilon_q}$ are vectors and $|\boldsymbol{\varepsilon_q}| > 0$, $\boldsymbol{\varepsilon_q}$ is collinear with $\boldsymbol{r_q}$ and shares the same direction.

Let $p(\boldsymbol{z} + \boldsymbol{R} \in S|\boldsymbol{z}_q)$ denote the probability of an n-dimensional ball, centered at $\boldsymbol{z}_q$ ($\boldsymbol{z} = \boldsymbol{z}_q$) with an error radius vector $\boldsymbol{R}$, being contained within the n-dimensional convex hull S. $\boldsymbol{z}$ and $\boldsymbol{R}$ are coordinate point and vector.

$$
\begin{cases}
p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_q) = p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_e) = p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_{lerp}) \\
= 1 \qquad\qquad\qquad\qquad\qquad\qquad |\boldsymbol{R}| \le |\boldsymbol{r_e}| < |\boldsymbol{r_q}| \\
\\
\exists \lambda \in [0,k), k<1 \quad \text{s.t.} p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_q) = p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_{lerp}) \\
= 1 > p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_e) \qquad\qquad |\boldsymbol{r_e}| < |\boldsymbol{R}| \le |\boldsymbol{r_q}| \\
\\
\exists \lambda \in (j,k), j>0, k<1 \quad \text{s.t.} |\boldsymbol{r_e}| < |\boldsymbol{r_q}| < |\boldsymbol{r_{lerp}}|, \\
p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_e) < p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_q) < p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_{lerp}) \\
\qquad\qquad\qquad\qquad\qquad\qquad |\boldsymbol{r_e}| < |\boldsymbol{r_q}| < |\boldsymbol{R}|
\end{cases}
$$

If $|\boldsymbol{r_e}| > |\boldsymbol{r_q}|$, it can be deduced in a similar manner.

If $|\boldsymbol{r_e}| = |\boldsymbol{r_q}|$ and both $\boldsymbol{z}_e$ and $\boldsymbol{z}_q$ approach the same boundary.

Then $p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_e) \le p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_{lerp}), \quad p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_q) \le p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_{lerp})$.

If $|\boldsymbol{r_e}| = |\boldsymbol{r_q}|$ and $\boldsymbol{z}_e$ and $\boldsymbol{z}_q$ approach different boundaries.

Then, it is analogous to the case of $|\boldsymbol{r_e}| < |\boldsymbol{r_q}| < |\boldsymbol{R}|$.

In conclusion, $p(\boldsymbol{z}+\boldsymbol{R}\in S|\boldsymbol{z}_{lerp})$ has a higher probability of still belonging to the convex hull S. □

Remark: Based on the above proof, we can rigorously conclude that employing interpolated points $\boldsymbol{z}_{lerp}$ yields superior robustness, consequently enhancing restoration performance.

To visually demonstrate how our method enhances VQ-LDM's robustness, we conduct a simple experiment using Monte Carlo sampling. We repeatedly generate reconstruction mapping points through the diffusion model on the same image, where numerical errors necessitate requantizing these points via vector quantization to align with the codebook. The standard deviation of corrected mapping points (pixel values in latent space) serves as our robustness metric. For this experiment, we employ the VQ-LDM-4 architecture with an 8,192-entry codebook and 3 latent channels. Results are presented in Figure 3. As clearly demonstrated in Figure 3, the interpolation-based latent initial state optimization significantly reduces the standard deviation, indicating more stable pixel-wise fluctuations in the generated latent channels and consequently higher robustness, properties that directly contribute to improved restoration performance. The Chebyshev center-based method achieves even greater reduction in standard deviation, empirically validating our earlier theoretical claim that it represents an elegant geometric solution.

Algorithm 1: Chebyshev center-based latent initial state optimization

Require: The embedding weights of codebook in VQ, denoted as $W^{K\times D}$, where K represents the size of the embedding space and D denotes the dimension of the embedding space ($K = 8192, D = 3$ in VQ-LDM-4). Decimal precision r.

Ensure: A dictionary mapping embedding mapping points $\boldsymbol{z}_q$ as keys to their corresponding Chebyshev centers as values.

Compute the point set s_1 that encloses the set of embedding mapping points (the set of codebook mapping points) s_2. Compute the Voronoi region of the union of s_1 and s_2.
for point p in $s_1 \bigcup s_2$ **do**
    Compute the convex hull boundaries of the region corresponding to point p, forming the boundary set B.
    for boundary equation e in B **do**
        Convert equation e into a linear inequality.
    end for
    Solve the linear programming problem formed by the system of inequalities and obtain the Chebyshev center c corresponding to point p.
    for value v in p **do**
        Reduce v with a decimal precision reduction of r based on the computing precision of the graphics card.
    end for
    Save the pair (p as key, c as value) into a dictionary.
end for

Experiments

We train and test VQ-LDM-4 with latent initial state optimization on nine standard benchmarks involving image restoration and video restoration. These nine datasets are CSD (Chen et al. 2021), Outdoor-Rain (Li, Cheong, and Tan 2019), RainDrop (Qian et al. 2018), RESIDE (Li et al. 2018), SIDD (Abdelhamed, Lin, and Brown 2018), GoPro (Nah, Hyun Kim, and Mu Lee 2017), NTURain (Chen et al. 2018), DAVIS (Khoreva, Rohrbach, and Schiele 2019) and RainSynAll100 (Yang et al. 2021). Regarding the training implementation, we mainly use two NVIDIA GeForce RTX 3090 GPUs, with a batch size of 8 and a fixed learning rate of 0.0001. We utilize image cropping techniques on the dataset, and randomly extract 16 patches of size 128×128 from each image. In the experiments, we set λ to 0.5 for interpolation-based latent initial state optimization with interpolation. However, as described in the preceding section (Table 1), if using interpolation as the latent initial state

ResShift		original results	λ									
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Super Resolution	PSNR	23.0160	23.0215	23.0262	23.0299	23.0327	23.0346	23.0355	**23.0356**	23.0348	23.0332	23.0306
	SSIM	0.6178	0.6181	0.6183	0.6185	0.6187	0.6189	0.6190	0.6191	**0.6192**	**0.6192**	0.6191
Image Inpainting	PSNR	17.8590	17.8598	17.8623	17.8676	17.8751	17.8847	17.8954	17.9070	17.9187	**17.9295**	17.9292
	SSIM	0.6624	0.6625	0.6627	0.6630	0.6632	0.6635	0.6637	0.6639	0.6641	**0.6642**	0.6641
Blind Face Restoration	PSNR	21.7634	**21.8434**	21.7158	21.3575	20.6089	19.9981	18.9254	18.0122	17.1126	16.2485	15.4786
	SSIM	0.6033	**0.6072**	0.6040	0.5925	0.5656	0.5445	0.5040	0.4710	0.4374	0.4054	0.3774

Table 1: The results of extending interpolation-based latent initial state optimization to ResShift. **Bold** represent the best results.

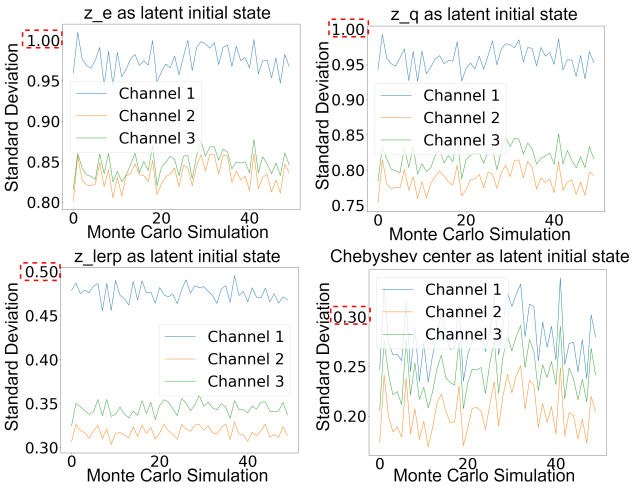

Figure 3: The standard deviation of results generated from multiple samplings on the same image with different latent initial states.

Restoration Performance		VQ-LDM-4		latent initial state Optimization	
		z_e	z_q	z_{lerp}	Chebyshev center
		PSNR/SSIM			
Image	CSD	24.75 0.8619	25.96 0.8675	27.80 0.8798	**28.16** **0.8813**
	Outdoor-Rain	24.21 0.7352	24.06 0.7328	25.01 0.7400	**25.13** **0.7431**
	RainDrop	23.90 0.7917	24.33 0.8113	25.38 0.7950	**25.41** **0.8132**
	RESIDE	23.46 0.9003	23.82 0.9138	24.25 0.9054	**25.02** **0.9178**
	SIDD	35.97 0.9193	36.44 0.9251	36.57 0.9273	**37.21** **0.9300**
Video	GoPro	26.16 0.7990	26.15 0.7987	26.20 0.7991	**26.53** **0.8013**
	NTURain	18.75 0.4890	17.58 0.4476	26.96 0.8415	**26.97** **0.8423**
	RainSynAll100	18.77 0.6812	18.06 0.6733	19.26 0.7074	**21.58** **0.7312**
	DAVIS	25.99 0.6511	26.44 0.6862	26.58 0.7352	**27.94** **0.7498**

Table 2: Comparison results of the performance at different latent initial state in VQ-LDM-4 with different image/video restoration datasets. **Bold** represent the best results.

optimization and aiming for improved restoration performance, a detailed analysis of the specific construction of each Voronoi subregion would be required. To ensure a fair comparison, VQ-LDM-4 with and without latent initial state optimization uses identical training epochs and number of function evaluations (NFEs).

The experiments encompass both image and video restoration tasks. While existing video restoration methods typically employ multi-frame inputs to recover the current frame (leveraging temporal information), our study focuses on investigating and addressing inherent flaws in VQ-LDM. The introduction of temporal information would interfere with our ability to diagnose VQ-LDM's core issues. Therefore, for video restoration experiments, we maintain the same frame-by-frame processing approach as in image restoration, deliberately avoiding the use of multi-frame inputs for current frame reconstruction.

Quantitative & Qualitative Results

The experimental results of VQ-LDM-4 with different latent initial states are presented in Figure 4 and Table 2. As shown in Table 2, using z_{lerp} as the latent initial state for training yields improvements across multiple datasets, with significantly greater performance gains compared to the train-free results (Table 1). However, our fixed $\lambda = 0.5$ setting may not fully realize the optimal potential of interpolation-based optimization. For instance, on the RainDrop and RESIDE datasets, while PSNR improves, SSIM shows slight degradation.

The Chebyshev center-based approach resolves this limitation by eliminating the need for manual λ tuning, achieving optimal performance across all datasets, making it the superior latent initial state choice for VQ-LDM. Visual analysis of the magnified regions in Figure 4 further confirms that both optimization methods enhance VQ-LDM's restoration quality, with the Chebyshev center-based method demonstrating notably more pronounced improvements.

Specifically, in quantitative results, interpolation-based latent initial state optimization achieved average improvements of 1.45 (PSNR) and 0.0469 (SSIM) across nine datasets compared to the best original VQ-LDM-4 results. For PSNR, the most significant improvement was observed on NTURain dataset (8.21), while the smallest gain was on GoPro dataset (0.04). Regarding SSIM, the maximum improvement reached 0.3525 on NTURain, while slight decreases of 0.0163 and 0.0084 occurred on RainDrop and

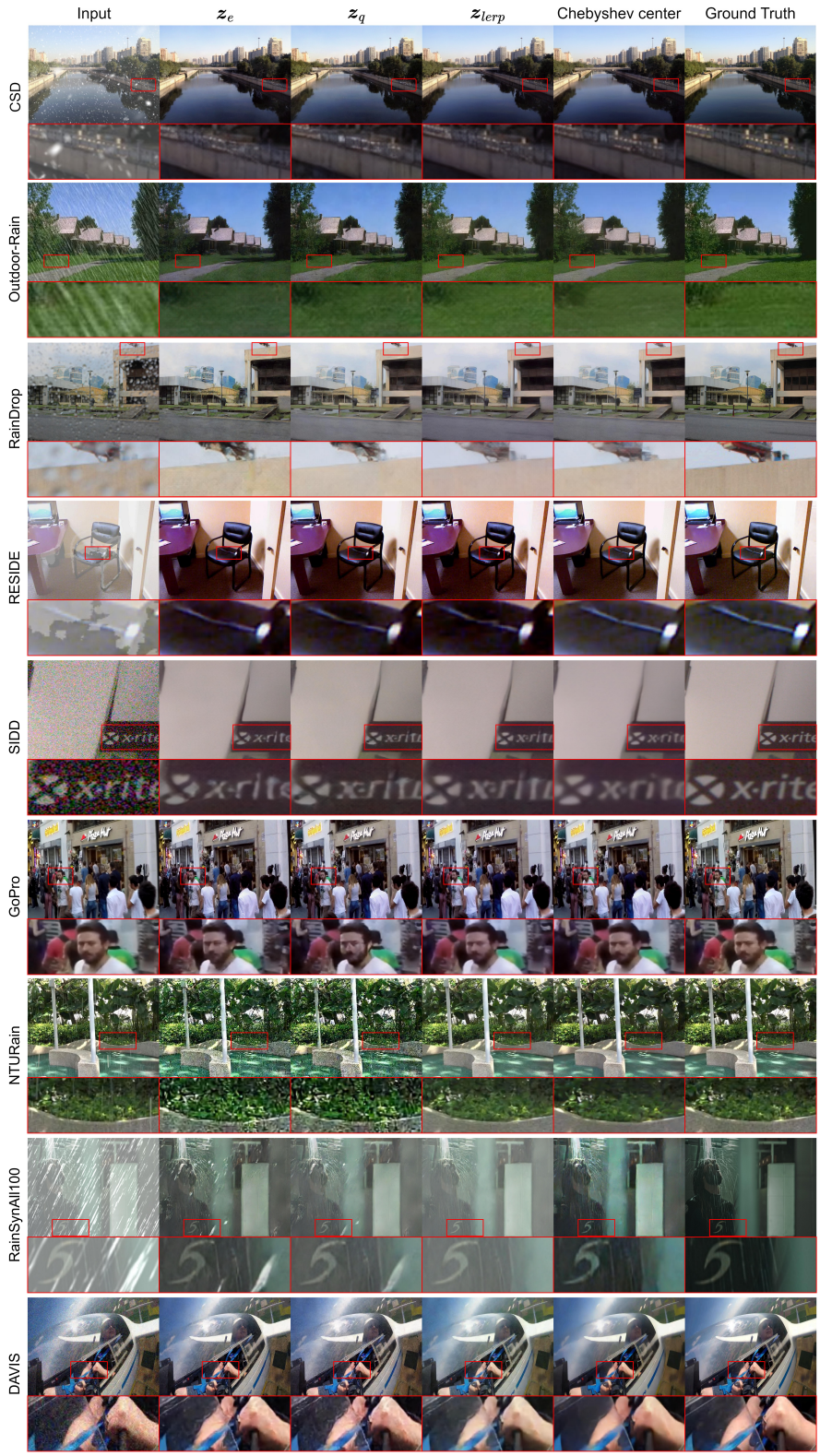

Figure 4: The imaging results obtained by diffusion models with different latent initial state on different image/video restoration datasets.

Performance Gain		z_{lerp}	Chebyshev center
		PSNR/SSIM	
Image	CSD	+1.84 +0.0123	+2.20 +0.0138
	Outdoor-Rain	+0.80 +0.0048	+0.92 +0.0079
	RainDrop	+1.05 -0.0163	+1.08 +0.0019
	RESIDE	+0.43 -0.0084	+1.20 +0.0040
	SIDD	+0.13 +0.0022	+0.77 +0.0049
Video	GoPro	+0.04 +0.001	+0.37 +0.0023
	NTURain	+8.21 +0.3525	+8.22 +0.3533
	RainSynAll100	+0.49 +0.0262	+2.81 +0.0500
	DAVIS	+0.14 +0.0490	+1.50 +00636

Table 3: The performance gains achieved by latent initial state optimization for the optimal VQ-LDM-4 across different datasets.

RESIDE respectively, as previously explained. Chebyshev center-based latent initial state optimization achieved average improvements of 2.11 (PSNR) and 0.0557 (SSIM) across nine datasets compared to the best original VQ-LDM-4 results. For PSNR, the maximum improvement reached 8.22 on NTURain dataset, while the minimum gain was 0.37 on GoPro dataset. Regarding SSIM, the highest improvement was 0.3533 on NTURain, with the smallest gain of 0.0019 on RainDrop. All performance gains are presented in Table 3.

In qualitative results, both methods enhance restored details to some extent, such as the gaps in stone railings (CSD dataset), the contours of tower cranes (RainDrop dataset), the clarity of text (SIDD dataset), and the shape of human arms (DAVIS dataset). On the other hand, the Chebyshev center-based method achieves superior image restoration, evidenced by more accurate chair color brightness (RESIDE dataset) and better removal of rain and haze degradation (DAVIS dataset).

This section extensively validates the impact of different latent initial states on the VQ-LDM architecture in restoration tasks, demonstrating that optimizing latent initial states enhances its recovery performance. Combined with the theoretical analysis from the previous section, these results robustly establish latent initial state optimization as an effective and well-grounded methodology.

Limitations & Suggestion

The experimental results in the preceding section demonstrate the superior performance of Chebyshev center-based latent initial state optimization over its interpolation-based counterpart. This naturally raises the question: Can Chebyshev center-based optimization completely replace interpolation-based optimization? Below, we discuss the inherent limitations of each approach and provide practical recommendations for their application.

The interpolation-based latent initial state optimization faces a key limitation: the optimal λ depends on the specific Voronoi subregion characteristics of each codebook entry, meaning distinct $\boldsymbol{z}_e/\boldsymbol{z}_q$ pairs require different optimal λ values. However, determining pixel-wise optimal λ values presents significant engineering challenges. For Chebyshev center-based optimization, the primary constraint lies in its dictionary lookup operation. As this CPU-intensive process necessitates frequent GPU-CPU switching during training, our experiments estimate an approximate $2\times$ increase in total training time.

We recommend using interpolation-based latent initial state optimization when low-cost performance improvement is desired for VQ-LDM in image restoration tasks, as it requires minimal code modifications and its feasibility and scalability have been demonstrated through train-free experiments. For scenarios prioritizing maximum performance gains without training time constraints, Chebyshev center-based latent initial state optimization is recommended.

Conclusion & Prospects

This paper investigates the flaws of VQ-LDM in image restoration tasks and provides analysis and solutions from a geometric perspective. Through experiments, we find that the main issue of VQ-LDM in image restoration lies in the lack of robustness in the codebook's mapping points. This causes the reconstruction mapping points generated by the diffusion model under error conditions to be requantized into incorrect Voronoi subregions, thereby degrading restoration performance. To address this, we propose latent initial state optimization. First, we introduce interpolation-based latent initial state optimization, which simply replaces the original mapping points through interpolation, and validate its effectiveness through train-free experiments and theoretical analysis. Subsequently, we further propose Chebyshev center-based latent initial state optimization, which uses linear programming to compute the Chebyshev center within Voronoi subregions and replaces the original mapping points to enhance robustness and restoration performance, this method represents an elegant theoretical solution from a geometric perspective. Extensive experiments demonstrate that latent initial state optimization is effective for VQ-LDM in image restoration tasks and can be applied to subsequent image restoration methods using the VQ-LDM architecture. For the two proposed latent initial state optimization methods, we also provide some prospects and considerations. For interpolation-based latent initial state optimization, one could find the k nearest mapping points to $\boldsymbol{z}_e$ and adaptively adjust λ through certain algorithms to obtain a relatively optimal λ. As for Chebyshev center-based latent initial state optimization, the focus would be more on engineering solutions to reduce the latency of dictionary lookup operations, thereby making it more hardware-friendly.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022ZD0161800), the National Natural Science Foundation of China under Grant 62271203, AI-Empowered Research Paradigm Reform and Discipline Leap Plan under Grant 2024AI01012 and the Open Research Fund of KLATASDS-MOE, ECNU. Besides, this work was supported by Bestpay AI Lab. Thanks for the reviewer's improvement suggestions. As an appendix cannot be added, if readers need the code for Algorithm 1, please contact us directly 52265901002@stu.ecnu.edu.cn.

References

Abdelhamed, A.; Lin, S.; and Brown, M. S. 2018. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1692–1700.

Chen, J.; Tan, C.-H.; Hou, J.; Chau, L.-P.; and Li, H. 2018. Robust video content alignment and compensation for rain removal in a cnn framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6286–6295.

Chen, W.-T.; Fang, H.-Y.; Hsieh, C.-L.; Tsai, C.-C.; Chen, I.; Ding, J.-J.; Kuo, S.-Y.; et al. 2021. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4196–4205.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*.

Khoreva, A.; Rohrbach, A.; and Schiele, B. 2019. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, 123–141. Springer.

Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; and Wang, Z. 2018. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1): 492–505.

Li, R.; Cheong, L.-F.; and Tan, R. T. 2019. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1633–1642.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Fei, B.; Dai, B.; Ouyang, W.; Qiao, Y.; and Dong, C. 2023. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*.

Luo, Z.; Gustafsson, F. K.; Zhao, Z.; Sjölund, J.; and Schön, T. B. 2023. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3883–3891.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Özdenizci, O.; and Legenstein, R. 2023. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Qian, R.; Tan, R. T.; Yang, W.; Su, J.; and Liu, J. 2018. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2482–2491.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4713–4726.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sun, H.; Li, W.; Liu, J.; Chen, H.; Pei, R.; Zou, X.; Yan, Y.; and Yang, Y. 2024. Coser: Bridging image and language for cognitive super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25868–25878.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *arXiv preprint arXiv:2305.07015*.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2023. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*.

Yang, W.; Tan, R. T.; Feng, J.; Wang, S.; Cheng, B.; and Liu, J. 2021. Recurrent multi-frame deraining: Combining physics guidance and adversarial learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8569–8586.

Yinhuai, W.; Jiwen, Y.; and Jian, Z. 2022. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 3.

Yue, Z.; Wang, J.; and Loy, C. C. 2023. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36.