Chain-of-Zoom: Extreme Super-Resolution via Scale Autoregression and Preference Alignment

Bryan Sangwoo Kim **Jeongsol Kim** **Jong Chul Ye**
KAIST AI
{bryanswkim, jeongsol, jong.ye}@kaist.ac.kr

Figure 1: Extreme super-resolution of photorealistic images by CoZ with up to 64× magnification (top) and 256× magnification (bottom). Fine details such as textures on a wall, wrinkles on a flag, and leaf veins are clearly seen.

Abstract

Modern single-image super-resolution (SISR) models deliver photo-realistic results at the scale factors on which they are trained, but collapse when asked to magnify far beyond that regime. We address this scalability bottleneck with *Chain-of-Zoom (CoZ)*, a model-agnostic framework that factorizes SISR into an autoregressive chain of intermediate scale-states with multi-scale-aware prompts. CoZ repeatedly re-uses a backbone SR model, decomposing the conditional probability into tractable sub-problems to achieve extreme resolutions without additional training. Because visual cues diminish at high magnifications, we augment each zoom step with multi-scale-aware text prompts generated by a vision-language model (VLM). The prompt extractor itself is fine-tuned using Generalized Reward Policy Optimization (GRPO) with a critic VLM, aligning text guidance towards human preference. Experiments show that a standard $4\times$ diffusion SR model wrapped in CoZ attains beyond $256\times$ enlargement with high perceptual quality and fidelity. Project Page: `https://bryanswkim.github.io/chain-of-zoom/`.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-4181

1 Introduction

The field of generative modeling has witnessed remarkable progress, enabling the synthesis of highly realistic data across various modalities, including images, text, and audio. A key application benefiting from these advancements is single-image super-resolution (SISR), which aims to reconstruct high-resolution (HR) details from a low-resolution (LR) input image. Super-resolution is a problem of core interest for effectively bridging the gap between low-cost imaging sensors and high-fidelity visual information; its usages range from enhancing consumer photographs and legacy media to improving critical details in medical imaging, satellite surveillance, and scientific visualization [2, 29, 31, 41, 44]. The standard approach to SISR is based on the posterior probability distribution:

$$p(\boldsymbol{x}_H \mid \boldsymbol{x}_L) \tag{1}$$

where the goal is to sample a plausible HR image $\boldsymbol{x}_H$ for a given input LR image $\boldsymbol{x}_L$. However, the mapping from $\boldsymbol{x}_L$ to $\boldsymbol{x}_H$ is highly complex and fundamentally ill-posed: a single LR image can correspond to a multitude of plausible HR images. This makes directly modeling the distribution extremely challenging for large magnification factors, and early attempts relying on interpolation or regression often produced blurry results [11, 14, 20, 51]. Recent emergence of powerful generative models (*e.g.*, diffusion-based models) has led to significant advancement in this task, providing strong generative priors over natural images that enable the synthesis of realistic textures and details consistent with the low-resolution input.

Specifically, existing methods leveraging such generative priors largely fall into two categories. One line of work frames SR as an inverse problem, utilizing a pre-trained generative model as a prior during inference time to find a realistic HR image consistent with the LR input [6–9, 21, 22]. While such inverse problem-solving methods benefit from being training-free, they typically require lengthy iterative optimization or sampling processes at inference time to enforce data consistency (*i.e.*, ensuring the downsampled HR prediction matches the original LR input), making them computationally expensive. Another line of work aims to incorporate this data consistency directly into the model's training objective, thereby enabling much faster inference [32, 43, 48, 49, 55, 56]. Modern state-of-the-art models within this category are capable of producing high-quality super-resolved images, even in a single inference step [48, 56].

However, these fast, trained super-resolution models suffer from a significant limitation: they are inherently upper-bounded by their training configuration and tend to collapse when presented with inputs requiring magnification beyond what they were trained on [23, 25, 58]. This failure occurs because the model's internal representations and learned restoration functions are tightly coupled to the specific scale and degradation seen during training [35]. Applying it outside this domain violates its learned assumptions, leading to severe artifacts, blurry outputs, or a complete failure to generate meaningful high-frequency details [12, 14, 20]. This lack of robustness severely restricts the practical applicability of these otherwise powerful models, demanding new models to be trained when the desired magnification factor exceeds what can be currently provided, which is highly inefficient.

In this work, we therefore propose to solve a fundamental question: *How can we effectively utilize super-resolution models to explore much higher resolutions than they were originally trained for?* Solving this question is critical in that it addresses the practical need for flexible and arbitrary-scale super-resolution, allowing users to magnify images to desired levels without being constrained by model training specifics. Furthermore, training models for extremely high magnification factors (*e.g.*, 16x, 32x) directly is often computationally prohibitive due to memory and time constraints [46]. Enabling the extension of existing, well-trained models (*e.g.*, 4x SR models) to higher factors offers a significantly more resource-efficient pathway to achieving extreme resolutions.

To address these fundamental challenges, we present *Chain-of-Zoom (CoZ)*, a novel framework for achieving extreme-resolution image generation beyond the training configurations of conventional super-resolution models. Specifically, we introduce intermediate scale-state modeling to bridge the gap between a low-resolution (LR) input and a high-resolution (HR) target image. These intermediate scale-states enable the decomposition of the conditional distribution in Eq. (1) into a series of tractable components, forming the basis of a scale-level autoregressive (AR) framework. Within this framework, models can progressively generate high-quality images at resolutions previously considered unattainable. In particular, building on the scale-level AR-2 model, we further propose a multi-scale-aware prompt extraction technique. This approach leverages Vision-Language Models (VLMs) to extract descriptive text prompts by attending to *multiple* scale-states throughout the

https://doi.org/10.52202/085713-4181

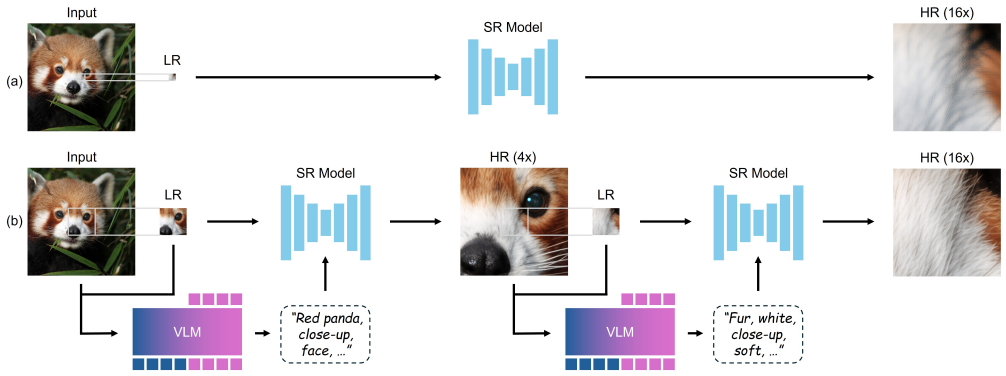


Figure 2: **(a) Conventional SR.** When an SR backbone trained for a fixed up-scale factor (*e.g.*, 4×) is pushed to much larger magnifications beyond its training regime, blur and artifacts are produced. **(b) Chain-of-Zoom (ours).** Starting from an LR input, a pretrained VLM generates a descriptive prompt, which—together with the image—is fed to the same SR backbone to yield the next HR scale-state. This prompt-and-upscale cycle is repeated, allowing a single off-the-shelf model to climb to extreme resolutions (16×–256×) while preserving sharp detail and semantic fidelity.

zooming process, enabling semantically aligned and coherent super-resolution. This is from the observation that at extreme resolutions, conditioning provided by the original signal $\boldsymbol{x}_L$ becomes insufficient, thus leading to unreasonable hallucinations by the SR model in cases.

Furthermore, to obtain text prompts of even richer detail that aligns with human preference, we fine-tune the prompt-extraction VLM under a novel RLHF pipeline leveraging GRPO [34]. A core part of this pipeline is the utilization of a critic VLM to score the outputs of the prompt extraction VLM, thus guiding it to produce prompts more aligned to human preference. Incorporated into the CoZ framework, our final VLM model successfully guides the super-resolution process towards reasonable high-quality results.

In summary, our contributions are as follows:

- We present *Chain-of-Zoom*, a scale-level autoregressive framework that decomposes super-resolution into a sequence of intermediate scale-states and multi-scale-aware prompts, enabling any existing SR model to reach much higher magnifications without retraining.
- We propose a novel RL pipeline for tuning prompt-extraction VLMs with GRPO. This pipeline incorporates appropriate reward functions and a critic reward model to endue multi-scale aware reasoning capabilities to the prompt-extraction VLM.

2 Related Work

Multi-Scale Image Generation and Super-Resolution. Unconditional multi-scale generators synthesize ever-larger images by passing coarse outputs through successive refinement stages. Cascaded Diffusion Models [17] pioneer this coarse-to-fine pipeline, while AnyresGAN [3], Scalespace-GAN [47], Generative Powers of Ten [45], ZoomLDM [53], and Make-a-Cheap-Scaling [16] share weights across latent zoom levels to reach megapixel resolutions. Because they are generation-based, these methods do not enforce consistency with a given low-resolution input. For true SR, PULSE [27] searches a GAN latent space, and Zoomed In, Diffused Out [28] alternates diffusion denoising with explicit up-sampling, but both do not explore extreme resolutions as in this work.

Autoregressive Factorizations. Classic autoregressive models such as PixelCNN, PixelRNN [39, 40] and VAR [38] predict spatial tokens sequentially within a fixed resolution. Pixel Recursive SR [10] extends this to super-resolution by autoregressing over *pixels* after each enlargement—effective for small factors but computationally prohibitive at extreme scales. The proposed CoZ instead autoregresses over *scale-states*: we factorize $p(\boldsymbol{x}_H \mid \boldsymbol{x}_L)$ into a tractable sequence of intermediate zoom distributions, enabling arbitrarily high magnifications without retraining at every factor.

https://doi.org/10.52202/085713-4181

Diffusion-Based Super-Resolution. Diffusion models have become the de-facto approach for high-fidelity SISR. SR3 [32] first denoised noisy HR guesses into realistic outputs with diffusion models. StableSR [43] reuses a diffusion prior for faster convergence, and prompt-aware variants (*e.g.*, SeeSR [49], SUPIR [55]) add textual conditioning to bolster semantic faithfulness. OSEDiff [48] distills the multi-step chain into a one-step denoising. Because of its accuracy and efficiency, we adopt OSEDiff as the backbone SR module in our CoZ demonstrations. However, CoZ is model-agnostic: the same scaling strategy can wrap any existing text-guided diffusion (or non-diffusion) SR network.

RL for Vision–Language Guidance. Reinforcement learning with human feedback (RLHF) is now widely used to align VLM behaviour with user preference. Early vision-grounded efforts such as LLaVA-RLHF [36] and LLaVACritic [50] employ reward models or critic networks to refine image-conditioned dialogue. Generalized Reward Policy Optimization (GRPO) was introduced by Shao et al. [34] as a policy-space alternative to PPO [33]. GRPO has since been adopted in vision tasks outside SR: Seg-Zero [26] uses GRPO to train VLMs for open-set semantic segmentation, while MetaSpatial [30] applies it to 3-D spatial reasoning in virtual environments. Building on these precedents, we are the first to bring GRPO to prompt-extraction in super-resolution. Our pipeline fine-tunes a prompt-extraction VLM with a composite reward objective unexplored in prior SR work.

3 Chain-of-Zoom

3.1 Intermediate Scale-State Modeling

In the CoZ framework, we propose to bridge the gap between a target HR image $\boldsymbol{x}_H \in \mathbb{R}^{d_n}$ and an input LR image $\boldsymbol{x}_L \in \mathbb{R}^{d_0}$ by introducing intermediate scale-states $\boldsymbol{x}_i \in \mathbb{R}^{d_i}$. Suppose that an image generative process is modeled as a sequence $(\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ where $\boldsymbol{x}_0 := \boldsymbol{x}_L$, $\boldsymbol{x}_n := \boldsymbol{x}_H$, and consecutive states have dimension ratio s (i.e. $d_i = sd_{i-1}$) larger than 1. Under the Markov assumption, the joint distribution could be modeled as $p(\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_n) = p(\boldsymbol{x}_0)\prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$. However, if the model follows a Markov chain structure, relying solely on the transition probability $p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1})$ leads to loss of high-frequency details as n increases (see Fig. 3). Inspired by recent work in inverse problems [8, 21] that demonstrate the effectiveness of text embeddings in reducing the solution space and improving super-resolution between consecutive scales, we therefore introduce latent variables $\boldsymbol{c}_i$ through text embeddings. The text prompt extraction supplements information of the overall zoom process.

Importantly, to reduce hallucinations caused by incorrect text guidance across scale, we find that multi-scale aware text extraction is necessary by feeding $\boldsymbol{x}_{i-1}$ *and* the coarser state $\boldsymbol{x}_{i-2}$ in prompt generation, leading to the conditional probability for the prompt:

$$p_\phi(\boldsymbol{c}_i \mid \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}). \tag{2}$$

Therefore, instead of using the Markov assumption, we propose AR-2 modeling of the image generative process with multi-scale-aware prompts as latent variables:

$$p(\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_n) = p(\boldsymbol{x}_0, \boldsymbol{x}_1)\prod_{i=2}^{n} p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}), \tag{3}$$

$$p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}) = \int p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i)p(\boldsymbol{c}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2})d\boldsymbol{c}_i. \tag{4}$$

Then, the joint distribution of the sequence $(\boldsymbol{x}_0, \boldsymbol{c}_1, \boldsymbol{x}_1, ..., \boldsymbol{c}_n, \boldsymbol{x}_n)$ is expressed as follows:

Proposition 1. *Given a sequence of scale-states $\boldsymbol{x}_i$ that follows a AR-2 structure and latent variables $\boldsymbol{c}_i$ that satisfy Eq. (2), the joint distribution is expressed as*

$$p(\boldsymbol{x}_0, \boldsymbol{c}_1, \boldsymbol{x}_1, ..., \boldsymbol{c}_n, \boldsymbol{x}_n) = p(\boldsymbol{x}_0, \boldsymbol{c}_1, \boldsymbol{x}_1)\prod_{i=2}^{n} p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i)p(\boldsymbol{c}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}). \tag{5}$$

Now, our objective function is maximizing the likelihood of the entire joint distribution of $\boldsymbol{x}_i$ and $\boldsymbol{c}_i$. Taking the logarithm of Eq. (5), we get the objective function to be maximized:

$$\mathcal{L} = \underbrace{\log p(\boldsymbol{x}_0, \boldsymbol{c}_1, \boldsymbol{x}_1)}_{\mathcal{L}_{\text{init}}} + \underbrace{\sum_{i=2}^{n} \log p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i)}_{\mathcal{L}_{\text{SR}}} + \underbrace{\sum_{i=2}^{n} \log p(\boldsymbol{c}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2})}_{\mathcal{L}_{\text{VLM}}} \tag{6}$$

where $\mathcal{L}_{\text{init}}$ represents the initial super-resolution step.

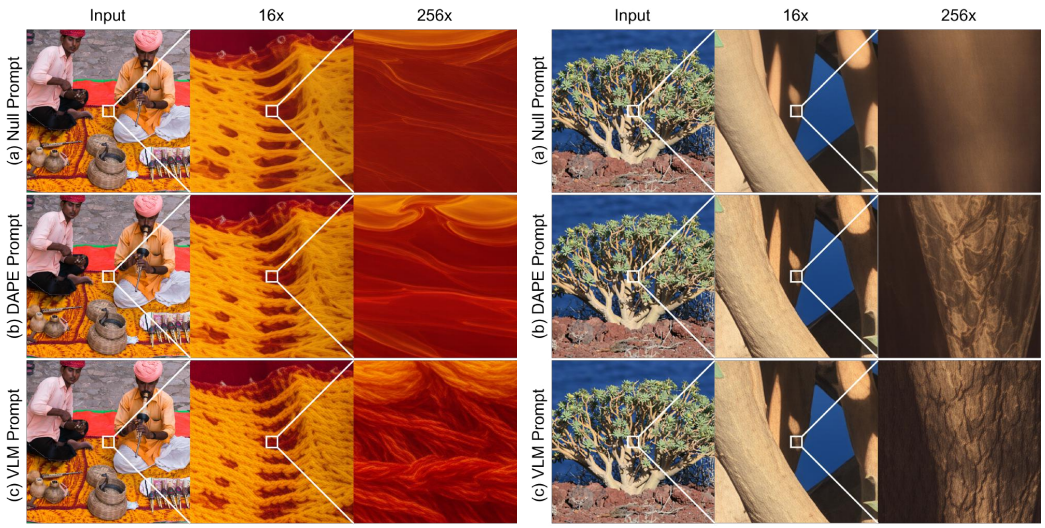


Figure 3: **Significance of proposed multi-scale-aware prompts:** *(a) Null prompt*: coarse structure is retained, but high-frequency details are smoothed out. *(b) DAPE prompt*: inserting text from a degradation-aware prompt extractor (DAPE) helps, yet the images lack intricate detail at large magnifications. *(c) VLM-generated prompts (ours)*: multi-scale prompts extracted by a VLM steer the SR backbone to synthesize realistic textures and crisp details.

3.2 Training Objective

The additive structure of the components in Eq. (6) allows for the independent optimization of each term. We achieve this via **next $\boldsymbol{x}_i$ prediction** and **next $\boldsymbol{c}_i$ prediction**, using parameterized models θ and ϕ, respectively.

Next $\boldsymbol{x}_i$ prediction. The training objective $\mathcal{L}_{\text{SR}}$ represents the likelihood of $\boldsymbol{x}_i$ given previous scale-states $\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}$ and description $\boldsymbol{c}_i$ for $\boldsymbol{x}_i$. Under the assumption that the distribution $p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i) := \mathcal{N}(\boldsymbol{x}_i; f_\theta(\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i), \sigma^2\mathbf{I})$ is Gaussian, where the parameterized model f_θ predicts the conditional mean of the distribution, the likelihood of $\boldsymbol{x}_i$ is equivalent to

$$\log p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i) = -\frac{1}{2\sigma^2}\|\boldsymbol{x}_i - f_\theta(\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i)\|^2 + C \tag{7}$$

where $C = -\frac{d_i}{2}\log(2\pi\sigma^2)$. To reduce the computational complexity of training f_θ, our key idea is that its dependency to $\boldsymbol{x}_{i-2}$ is only through the multi-scale-aware prompt, i.e. $\boldsymbol{c}_i = \boldsymbol{c}_i(\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2})$, leading to $f_\theta(\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i) = f_\theta(\boldsymbol{x}_{i-1}, \boldsymbol{c}_i(\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}))$. Maximizing the simplified likelihood thus reduces to minimizing the mean-squared error (MSE) between the predicted HR patch from $\boldsymbol{x}_{i-1}$ and the ground truth—precisely the loss most SR backbones are already trained with. In this work, we perform experiments with a backbone SR model trained via settings in Sec. 4.1, yet our framework is model-agnostic.

Next $\boldsymbol{c}_i$ prediction. Recall that the dependency to the $\boldsymbol{x}_{i-2}$ in AR-2 model is through the multi-scale aware prompt extraction, which supplements information of the overall zoom process and reduces hallucinations caused by incorrect text guidance. For a single zoom step *i*, the prompt $\boldsymbol{c}_i = (c_{i,1}, \cdots, c_{i,T_i})$ is a token sequence conditioned on the current and previous image, i.e. $\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}$. Modern VLMs model this distribution autoregressively:

$$p_\phi(\boldsymbol{c}_i \mid \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}) = \prod_{t=1}^{T_i} p_\phi(c_{i,t} \mid \boldsymbol{x}_{i-1}, , \boldsymbol{x}_{i-2}, c_{i,<t}) \tag{8}$$

where $c_{i,<t} = (c_{i,1}, \cdots, c_{i,t-1})$. Maximizing the log-likelihood $\log p_\phi(\boldsymbol{c}_i \mid \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2})$ therefore amounts to minimizing the negative log-likelihood (cross-entropy) for each token:

$$\mathcal{L}_{\text{VLM}}^{(i)} = -\log p_\phi(\boldsymbol{c}_i \mid \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}) = -\sum_{t=1}^{T_i} \log p_\phi(c_{i,t} \mid \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, c_{i,<t}). \tag{9}$$

https://doi.org/10.52202/085713-4181

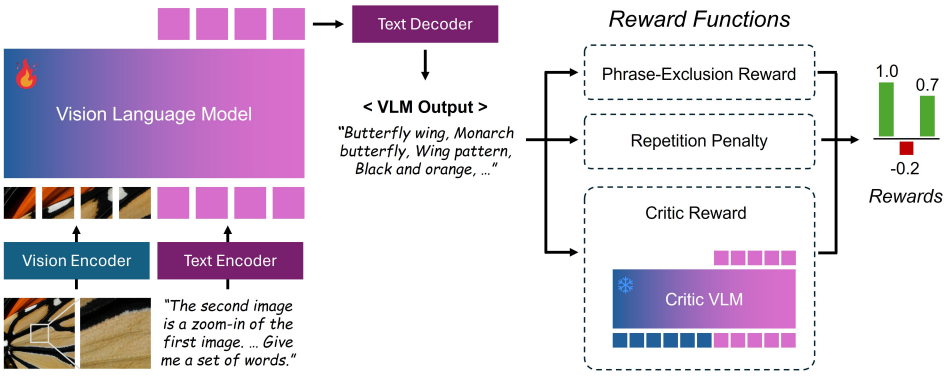


Figure 4: **GRPO Training Framework.** At every zoom step, multi-scale image crops are fed to the base VLM, which generates candidate prompts after perceiving input images. A critic VLM scores the prompt for semantic quality, while phrase-exclusion and repetition penalties enforce conciseness and relevance. The weighted sum of these rewards forms the GRPO signal that iteratively fine-tunes the base VLM, steering it towards prompts that best guide extreme-scale super-resolution.

Eq. (9) is exactly the standard next-token cross-entropy loss used to pre-train modern VLMs; hence our framework can employ any off-the-shelf VLM whose weights already maximize this objective.

Inference. Given pre-trained parameterized models θ and ϕ, the sequence ($\boldsymbol{x}_0$, $\boldsymbol{c}_1$, $\boldsymbol{x}_1$, ..., $\boldsymbol{c}_n$, $\boldsymbol{x}_n$) can be generated recursively. Starting from the low-resolution image $\boldsymbol{x}_L = \boldsymbol{x}_0$, a description for the next scale, $\boldsymbol{c}_1 \sim p_\phi(\boldsymbol{c}_1 \mid \boldsymbol{x}_0)$, is first sampled. Then, the next scale state is generated by sampling $\boldsymbol{x}_1 \sim p_\theta(\boldsymbol{x}_1 \mid \boldsymbol{x}_0, \boldsymbol{c}_1)$. For subsequent steps, the description at scale i is sampled as $\boldsymbol{c}_i \sim p_\phi(\boldsymbol{c}_i \mid \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2})$, followed by sampling the image at that scale as $\boldsymbol{x}_i \sim p_\theta(\boldsymbol{x}_i \mid \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i)$. This sequential sampling process generates specific, plausible high-resolution outputs $\boldsymbol{x}_n$ without needing to model the full marginal distribution $p(\boldsymbol{x}_0, ..., \boldsymbol{x}_n)$ explicitly. When using SR backbone models that require input and output dimensions to be identical (*e.g.*, Stable-diffusion-based SR models [43, 48, 49, 55]), a fixed-size window is cropped from the HR image and resized to the required dimension. Thus, super-resolution operates in local regions, and achieving outputs of entire images would require multiple runs of CoZ.

3.3 Training Multi-Scale-Aware Prompt Extraction using RL

At extreme magnification factors, the visual evidence in the input image becomes extremely sparse, causing the SR backbone model to rely more heavily on text prompts. To curb the ensuing drift towards implausible high-frequency hallucinations, we fine-tune the prompt-extraction VLM so that its textual guidance aligns with human aesthetic and semantic preferences. Our fine-tuning pipeline (Fig. 4) adopts Generalized Reward Policy Optimization (GRPO). For each zoom step i, the VLM receives multi-scale image crops $(\boldsymbol{x}_{i-2}, \boldsymbol{x}_{i-1})$ and produces a candidate prompt $\boldsymbol{c}_i$. The prompt is scored by a set of task-specific reward functions, and the weighted sum $R(\boldsymbol{c}_i)$ drives the GRPO update to align the VLM prompts with human preference. The overall reward $R(\boldsymbol{c}_i)$ is a weighted sum of three components, each targeting a distinct failure mode observed during preliminary experiments:

$$R(\boldsymbol{c}_i) = w_{\text{critic}}\, R_{\text{critic}} + w_{\text{phrase}}\, R_{\text{phrase}} + w_{\text{rep}}\, R_{\text{rep}} \tag{10}$$

Critic Preference Reward (R_{critic}). A stronger vision–language critic VLM judges the candidate prompt in the context of the input multi-scale image crops and assigns a raw score in $[0, 100]$. We linearly rescale this score to $[0, 1]$ and treat it as a proxy for human preference, thereby imbuing the prompt-extraction VLM with the critic VLM's higher-level semantic priors.

Phrase-Exclusion Reward (R_{phrase}). Multi-image conditioning occasionally leads the prompt-extraction VLM to emit viewpoint markers such as "*first image*" or "*second image*," which are meaningless to the downstream SR model. We therefore issue a reward of 1 if none of a predefined blacklist of such phrases appear, and 0 otherwise.

Repetition Penalty (R_{rep}). Following Yeo et al. [54], we compute the fraction of repeated n-grams in the prompt and give a negative reward (down to -1) for a higher repetition ratio.

https://doi.org/10.52202/085713-4181

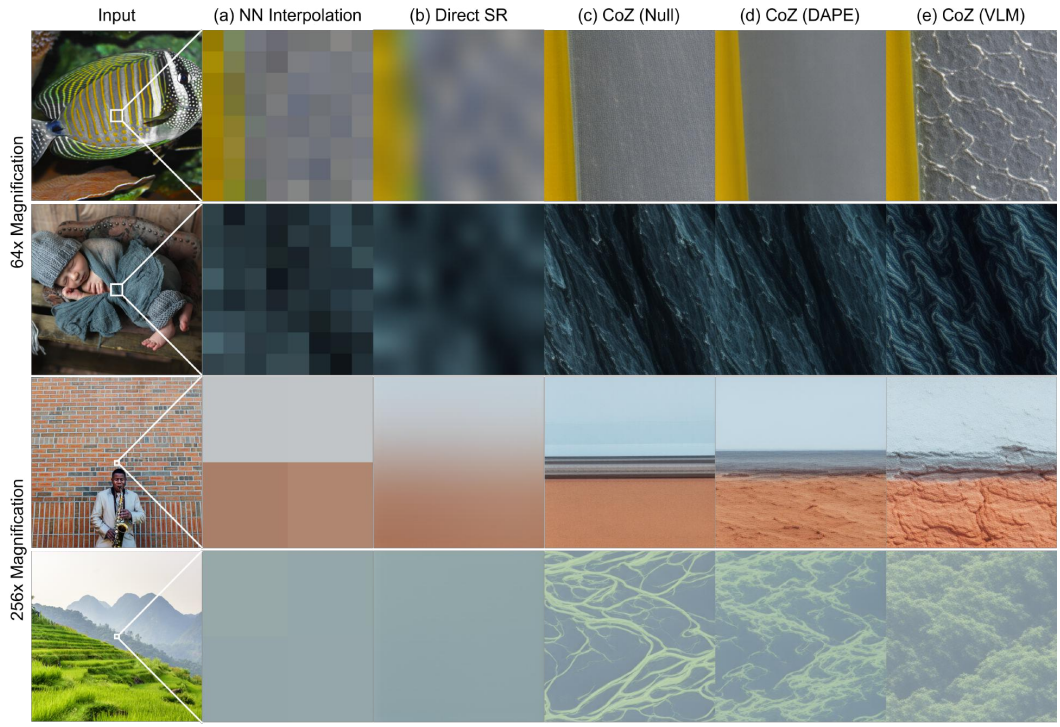


Figure 5: **Qualitative Results.** For each input image, super-resolution is performed on different magnifications with various methods: **(a) Nearest neighbor interpolation**; **(b) One-step direct SR** with the backbone SR model; **(c-e) Variants of CoZ** with different text prompts. The CoZ framework shows significantly better performance at large magnifications. Furthermore, using CoZ with VLM prompts assists the SR model in generating realistic details without hallucinations.

4 Experiments

4.1 Experimental Settings

We adopt the setup of prior work [49, 48] and train OSEDiff [48] as the backbone SR model with the LSDIR [24] dataset and 10K images from FFHQ [18]. We use Stable Diffusion 3.0 [13] as the backbone diffusion model and adopt a coarse-to-fine training strategy: first training on random degradation, and then training specifically for $4\times$ magnifications. Text guidance is provided by Degradation-Aware Prompt Extractor (DAPE) [49] as the naive prompt extractor, while Qwen2.5-VL-3B-Instruct [37] is used as the prompt-extraction VLM. RLHF training with GRPO is performed with InternVL2.5-8B [5] as the critic VLM. The same dataset used for training the backbone SR model is also used for GRPO training, and weights are given as: $w_{\text{critic}} = 1.0$, $w_{\text{phrase}} = 0.5$, $w_{\text{rep}} = 0.5$.

Evaluation is performed on the training datasets of DIV2K [1] and DIV8K [15], consisting of 800 images and 1500 images, respectively. Each image is resized and center-cropped to resolution of 512×512 to be input to the SR model. For four recursions, the HR image of the previous zoom is center-cropped and resized by a scale of 4 back to the resolution of 512×512.

4.2 Comparison Results

We perform comparison across four recursions for various methods. Specifically, we compare between nearest neighbor interpolation, direct magnification via one-step SR, and three versions of the proposed CoZ leveraging different prompts (*i.e.*, Null, DAPE, VLM).

Qualitative Comparison. Qualitative results in Fig. 5 show that nearest neighbor interpolation and one-step direct SR fall off at higher scales, while CoZ variants produce images of better quality. Thus, incorporating VLM prompts helps overcome the sparsity of the original input signal.

https://doi.org/10.52202/085713-4181

Table 1: Quantitative comparison on no-reference metrics. **Bold**: best, Underline: second-best.

Scale	Method	DIV2K NIQE↓	DIV2K MUSIQ↑	DIV2K MANIQA↑	DIV2K CLIPIQA↑	DIV8K NIQE↓	DIV8K MUSIQ↑	DIV8K MANIQA↑	DIV8K CLIPIQA↑
4×	NN Interpolation	12.1252	39.96	0.3396	0.2630	13.1984	40.26	0.3472	0.2672
	Direct SR	4.7320	67.00	0.6344	0.7005	4.8631	66.29	0.6359	0.6946
	CoZ (Null)	4.7706	66.99	0.6309	0.6977	4.9011	66.23	0.6325	0.6897
	CoZ (DAPE)	4.7312	67.01	0.6344	0.7004	4.8607	66.29	0.6359	0.6946
	CoZ (VLM)	**4.6572**	**67.10**	**0.6360**	**0.7017**	**4.8099**	**66.37**	**0.6370**	**0.6953**
16×	NN Interpolation	22.1215	24.01	0.3378	0.2346	22.2744	24.94	0.3465	0.2585
	Direct SR	7.2183	51.25	0.5406	0.6080	7.5855	50.17	0.5473	0.6035
	CoZ (Null)	6.5016	**59.19**	0.5859	**0.6686**	6.7898	**58.04**	0.5881	0.6618
	CoZ (DAPE)	6.5456	58.83	0.5946	0.6609	6.8607	57.79	0.5964	**0.6628**
	CoZ (VLM)	**6.3957**	58.81	**0.5970**	0.6574	**6.6500**	57.99	**0.6006**	0.6615
64×	NN Interpolation	27.4051	37.69	0.3803	0.3690	27.7533	37.13	0.3861	0.3837
	Direct SR	16.5915	22.54	0.3995	0.4309	16.5874	22.97	0.4069	0.4451
	CoZ (Null)	8.3500	51.82	0.5627	0.6305	8.5694	50.96	0.5638	0.6240
	CoZ (DAPE)	8.6598	51.77	0.5726	0.6262	8.7669	50.40	0.5714	0.6274
	CoZ (VLM)	**8.2335**	**52.13**	**0.5788**	**0.6315**	**8.2992**	**51.20**	**0.5787**	**0.6282**
256×	NN Interpolation	34.8461	27.01	0.4179	0.5259	37.2612	26.98	0.4184	0.5299
	Direct SR	16.1749	28.89	0.4470	0.5196	15.8667	28.90	0.4464	0.5256
	CoZ (Null)	10.0456	46.28	0.5510	0.5857	10.0630	46.56	0.5479	0.5899
	CoZ (DAPE)	10.4569	46.22	0.5564	0.5889	10.2788	45.81	0.5535	0.5984
	CoZ (VLM)	**9.8260**	**47.83**	**0.5692**	**0.5986**	**9.6405**	**47.25**	**0.5646**	**0.6041**

Quantitative Comparison. Quantitative results are given in Tab. 1. Due to the non-availability of ground-truth images for $256\times$ magnifications, we follow [27] and evaluate performance on no-reference perceptual metrics. Specifically, we use the metrics NIQE [57], MUSIQ [19], MANIQA-pipal [52], CLIPIQA [42] for a thorough evaluation. At low scales (*i.e.*, Scale $4\times$), difference between methods is minimal, but at high scales (*i.e.*, Scales $64\times$, $256\times$) the proposed framework shows consistently better performance. Furthermore, prompts by DAPE show comparable performance at low scales but fall off at higher scales, while VLM-generated prompts exhibit significantly better performance, supporting our claim that prompt-extraction by VLMs make up for the deficient visual conditioning provided by the initial image.

Quantitative Comparison with Baseline Methods. In Tab. 2, we further provide quantitative comparison with baseline methods: arbitrary-scale SR methods (LIIF [4]) and direct super-resolution of diffusion-based SR methods (SeeSR [49], S3Diff [56]). All three baselines show greatly degraded performance at high magnifications. For the case of S3Diff, we additionally show quantitative results for applying CoZ to the pretrained, freely accessible S3Diff, leveraging our multi-scale aware GRPO fine-tuned VLM. All results clearly confirm the significant performance improvement by CoZ. Additional results for performing CoZ with OSEDiff leveraging the Stable Diffusion v2.1 backbone is provided in Appendix E.

4.3 GRPO for VLM

GRPO Training. Reward graphs for training the prompt-extraction VLM with different critic VLMs are shown in Fig. 6 and Fig. 7. When using InternVL2.5-8B [5] as the critic VLM, phrase exclusion reward and repetition penalty converge to 1.00 and 0.00 (respectively) in the early stages of training, while the critic reward increases gradually throughout the training process. Similar trends are observed when using Qwen2.5-VL-7B-Instruct [37] as the critic VLM, proving the robustness of our method. Additional quantitative results for this case is provided in Appendix F.

Preference Alignment. Using an off-the-shelf VLM for prompt-extraction can cause unwanted hallucinations to occur in the zoom process. An example case is shown in Fig. 8 (Top), where the off-the-shelf VLM generates improper prompts due to insufficient knowledge of the initial image at high magnifications. By inducing the VLM to generate multi-scale-aware prompts by conditioning on $(\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2})$, we can produce more suitable prompts Fig. 8 (Middle). Finally, using the VLM fine-tuned with GRPO we can produce high-quality samples while reducing unwanted hallucinations as in Fig. 8 (Bottom). We further prove that the VLM after undergoing GRPO training is better aligned with human preference through user study. For this, we follow prior work [27], and perform a MOS (mean-opinion-score) test on various samples. Results and details are included in Appendix D.

https://doi.org/10.52202/085713-4181

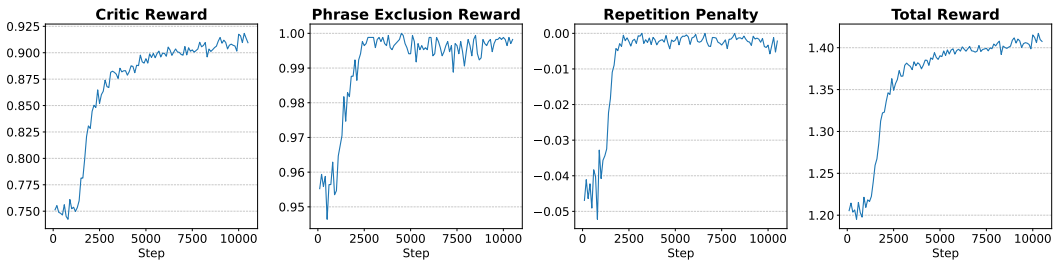


Figure 6: Reward graphs of using **InternVL2.5-8B** as the critic VLM, evaluated on a validation set. Values for Critic Reward, Phrase Exclusion Reward, Repetition Penalty, and Total Reward increase throughout the training process.

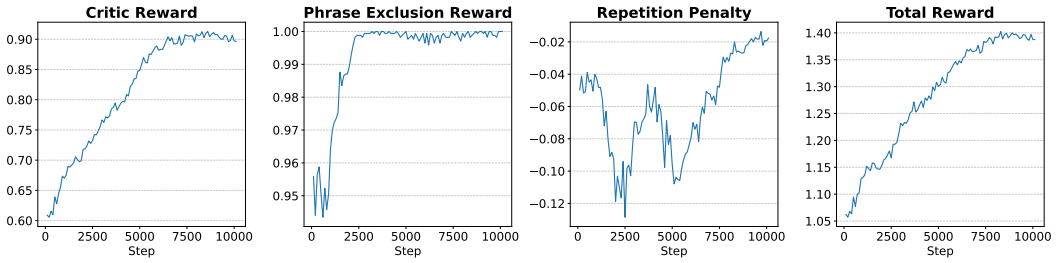


Figure 7: Reward graphs of using **Qwen2.5-VL-7B-Instruct** as the critic VLM, evaluated on a validation set. Values for Critic Reward, Phrase Exclusion Reward, Repetition Penalty, and Total Reward increase throughout the training process.

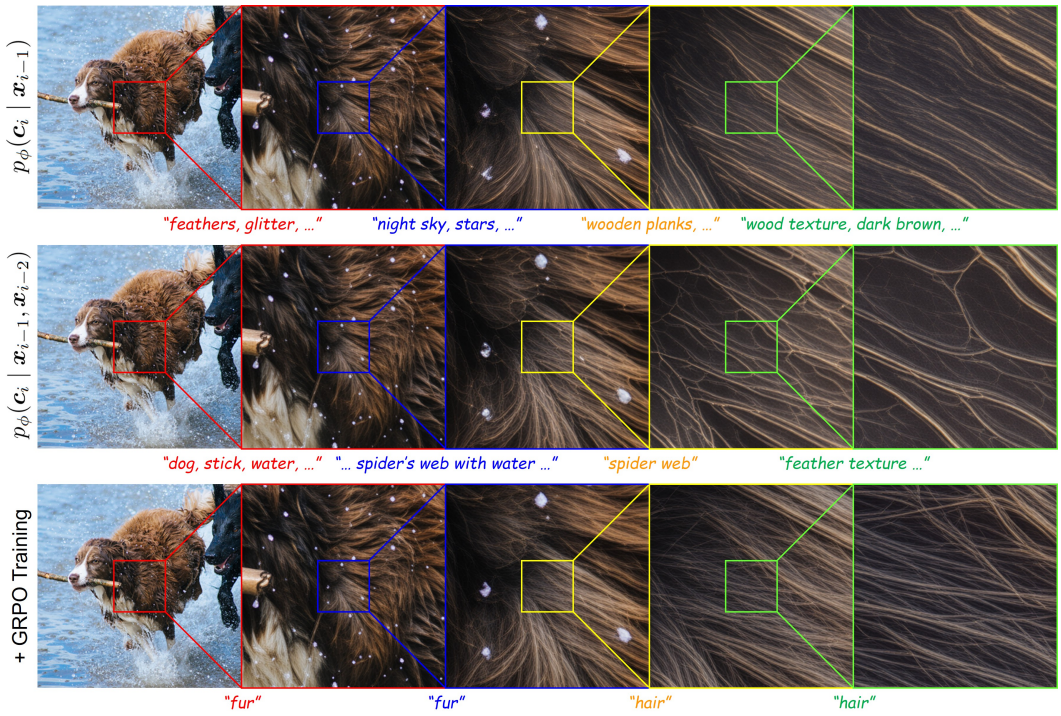


Figure 8: RLHF training with GRPO assists the prompt-extraction VLM in creating meaningful prompts for accurate guidance. (Top) **Base VLM**: generating prompts only from the LR input causes unwanted hallucinations as shown by the incorrect prompts; (Middle) **Multi-scale image prompts** are helpful at low scales (*e.g.*, accurate prompt of "dog, stick, water, ...") but fail at high scales; (Bottom) **VLM aligned with human preference** guides samples with improved text guidance.

https://doi.org/10.52202/085713-4181

Table 2: Quantitative comparison with baseline methods. **Best**, Second-Best.

Scale	Method	DIV2K NIQE↓	DIV2K MUSIQ↑	DIV2K MANIQA↑	DIV2K CLIPIQA↑	DIV8K NIQE↓	DIV8K MUSIQ↑	DIV8K MANIQA↑	DIV8K CLIPIQA↑
4×	LIIF	6.6210	57.47	0.5456	0.4587	6.8050	55.52	0.5386	0.4545
	SeeSR	5.5208	57.56	0.5387	0.5535	5.5940	55.50	0.5346	0.5385
	S3Diff	4.9803	65.82	0.6361	0.6596	5.0305	64.08	0.6328	0.6481
	S3Diff + CoZ	**4.8637**	**67.18**	**0.6459**	**0.6835**	**4.9414**	**65.31**	**0.6405**	**0.6680**
16×	LIIF	11.4815	25.94	0.2860	0.3024	11.8734	26.73	0.2896	0.3178
	SeeSR	9.6798	41.68	0.4310	0.4669	9.8865	40.02	0.4210	0.4601
	S3Diff	6.7383	51.14	0.5305	0.5886	7.2399	50.32	0.5370	0.5841
	S3Diff + CoZ	**6.7310**	**56.82**	**0.5752**	**0.6168**	**6.8698**	**56.46**	**0.5836**	**0.6218**
64×	LIIF	16.9215	20.02	0.3451	0.4131	17.4912	20.43	0.3522	0.4250
	SeeSR	16.9095	21.83	0.4106	0.4193	16.9275	22.68	0.4167	0.4309
	S3Diff	16.1421	21.54	0.3893	0.4917	16.5506	22.00	0.3968	0.5089
	S3Diff + CoZ	**8.7770**	**48.90**	**0.5490**	**0.5801**	**8.9794**	**48.07**	**0.5560**	**0.5909**
256×	LIIF	23.8949	26.05	0.4108	0.5380	24.3300	26.01	0.4116	0.5423
	SeeSR	20.9635	25.81	0.4438	0.5193	19.9628	25.94	0.4429	0.5203
	S3Diff	16.7809	25.92	0.4324	0.5359	16.9952	25.91	0.4312	0.5415
	S3Diff + CoZ	**10.7668**	**43.59**	**0.5438**	**0.5570**	**10.5001**	**43.31**	**0.5417**	**0.5673**

Table 3: Runtime analysis for different methods (seconds).

Scale	Phase	Direct SR (DAPE)	CoZ (Null)	CoZ (DAPE)	CoZ (VLM)
4×	SR	0.1467	0.1443	0.1460	0.1445
	PE	0.0136	0.0000	0.0130	0.3777
16×	SR	0.1462	0.2886	0.2912	0.2896
	PE	0.0130	0.0000	0.0254	0.7068
64×	SR	0.1462	0.4329	0.4363	0.4349
	PE	0.0131	0.0000	0.0382	1.0301
256×	SR	0.1462	0.5774	0.5816	0.5802
	PE	0.0132	0.0000	0.0509	1.3505

4.4 Runtime Analysis

Runtime analysis for direct super-resolution and CoZ variants across varying scales (4× to 256×) is given in Tab. 3. The average inference time required (in seconds) to apply CoZ on a single image is evaluated on 500 images of the DIV2K dataset. We divide inference into two phases: super-resolution (SR) and prompt extraction (PE); computational time required for each phase is analyzed accordingly.

5 Conclusion

This paper tackles the long-standing scalability gap in single-image super-resolution: state-of-the-art models excel at their trained scale factors yet fail when asked to enlarge images far beyond that range. Specifically, we introduced *Chain-of-Zoom (CoZ)*, a scale-level autoregressive framework that transforms any existing SR backbone into an extreme-magnification engine by decomposing the LR to HR mapping into a sequence of intermediate scale-states and multi-scale-aware prompts. CoZ is model-agnostic, requires no retraining of the base network, and thus offers a cost-effective path up to extreme resolutions. In particular, to maintain semantic coherence as visual evidence thins out, we leverage a multi-scale-aware prompt extractor driven by a VLM fine-tuned through a GRPO-based RLHF pipeline. Overall, CoZ yields sharp, realistic results at extreme scales while keeping inference efficient. By decoupling super-resolution performance from fixed training magnifications and demonstrating the value of aligned textual guidance, our work opens new avenues for resource-frugal image enhancement and lays a foundation for future exploration of learned zoom policies, domain-specific reward functions, and adaptive backbone selection.

Limitation and Potential Negative Impacts. While CoZ enables extreme super-resolution with high visual fidelity, it requires repeated application for extreme magnification, which may cause error accumulation over iterations. Moreover, high-fidelity generation from low-resolution inputs may raise concern regarding misinformation or unauthorized reconstruction of sensitive visual data.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea under Grant RS-2024-00336454, and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2024-00457882, AI Research Hub Project). This work was also supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2025-RS-2020-II201461).

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017.

[2] Eric Betzig, George H Patterson, Rachid Sougrat, O Wolf Lindwasser, Scott Olenych, Juan S Bonifacino, Michael W Davidson, Jennifer Lippincott-Schwartz, and Harald F Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *science*, 313(5793):1642–1645, 2006.

[3] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *European conference on computer vision*, pages 170–188. Springer, 2022.

[4] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021.

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[6] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=nJJjv0JDJju`.

[7] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=OnD9zGAGT0k`.

[8] Hyungjin Chung, Jong Chul Ye, Peyman Milanfar, and Mauricio Delbracio. Prompt-tuning latent diffusion models for inverse problems. *arXiv preprint arXiv:2310.01110*, 2023.

[9] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=DsEhqQtfAG`.

[10] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. Pixel recursive super resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 5439–5448, 2017.

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014.

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[14] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.

 https://doi.org/10.52202/085713-4181

[15] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3512–3516. IEEE, 2019.

[16] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In *European Conference on Computer Vision*, pages 39–55. Springer, 2024.

[17] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23 (47):1–33, 2022.

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[19] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

[20] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 2003.

[21] Jeongsol Kim, Geon Yeong Park, Hyungjin Chung, and Jong Chul Ye. Regularization by texts for latent diffusion inverse solvers. *arXiv preprint arXiv:2311.15658*, 2023.

[22] Jeongsol Kim, Bryan Sangwoo Kim, and Jong Chul Ye. Flowdps: Flow-driven posterior sampling for inverse problems. *arXiv preprint arXiv:2503.08136*, 2025.

[23] Christian Ledig, Lucas Theis, Ferenc Husz'ar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[24] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023.

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[26] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

[27] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020.

[28] Brian B Moser, Stanislav Frolov, Tobias C Nauen, Federico Raue, and Andreas Dengel. Zoomed in, diffused out: Towards local degradation-aware multi-diffusion for extreme image super-resolution. *arXiv preprint arXiv:2411.12072*, 2024.

[29] Ozan Oktay, Wenjia Bai, Matthew Lee, Ricardo Guerrero, Konstantinos Kamnitsas, Jose Caballero, Antonio de Marvao, Stuart Cook, Declan O'Regan, and Daniel Rueckert. Multi-input cardiac image super-resolution using convolutional neural networks. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19*, pages 246–254. Springer, 2016.

[30] Zhenyu Pan and Han Liu. Metaspatial: Reinforcing 3d spatial reasoning in vlms for the metaverse. *arXiv preprint arXiv:2503.18470*, 2025.

[31] Saiprasad Ravishankar and Yoram Bresler. MR image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE transactions on medical imaging*, 30(5):1028–1041, 2010.

[32] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021.

[33] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[34] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[35] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018.

[36] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[37] Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

[38] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

[39] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

[40] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016.

[41] Lena Wagner, Lukas Liebel, and Marco Körner. Deep residual learning for single-image super-resolution of multi-spectral satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:189–196, 2019.

[42] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2555–2563, 2023.

[43] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12): 5929–5949, 2024.

[44] Peijuan Wang, Bulent Bayram, and Elif Sertel. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232:104110, 2022.

[45] Xiaojuan Wang, Janne Kontkanen, Brian Curless, Steven M Seitz, Ira Kemelmacher-Shlizerman, Ben Mildenhall, Pratul Srinivasan, Dor Verbin, and Aleksander Holynski. Generative powers of ten. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7173–7182, 2024.

[46] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020.

[47] Krzysztof Wolski, Adarsh Djeacoumar, Alireza Javanmardi, Hans-Peter Seidel, Christian Theobalt, Guillaume Cordonnier, Karol Myszkowski, George Drettakis, Xingang Pan, and Thomas Leimkühler. Learning images across scales using adversarial training. *ACM Transactions on Graphics*, 43(4):131, 2024.

[48] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37:92529–92553, 2024.

[49] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024.

[50] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024.

[51] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

 https://doi.org/10.52202/085713-4181

[52] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022.

[53] Srikar Yellapragada, Alexandros Graikos, Kostas Triaridis, Prateek Prasanna, Rajarsi R Gupta, Joel Saltz, and Dimitris Samaras. Zoomldm: Latent diffusion model for multi-scale image generation. *arXiv preprint arXiv:2411.16969*, 2024.

[54] Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms, 2025. URL `https://arxiv.org/abs/2502.03373`.

[55] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024.

[56] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024.

[57] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

[58] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[59] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29733–29735, 2025.

A Proofs

Proposition 1. *Given a sequence of scale-states $\boldsymbol{x}_i$ that follows a AR-2 structure and latent variables $\boldsymbol{c}_i$ that satisfy Eq. (2), the joint distribution is expressed as*

$$p(\boldsymbol{x}_0, \boldsymbol{c}_1, \boldsymbol{x}_1, ..., \boldsymbol{c}_n, \boldsymbol{x}_n) = p(\boldsymbol{x}_0, \boldsymbol{c}_1, \boldsymbol{x}_1) \prod_{i=2}^{n} p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i) p(\boldsymbol{c}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}). \tag{5}$$

Proof. By substituting Eq. (4) to Eq. (3) and expanding the base term $p(\boldsymbol{x}_0, \boldsymbol{x}_1)$, we obtain:

$$\begin{aligned}
&\left[\int p(\boldsymbol{x}_0, \boldsymbol{c}_1, \boldsymbol{x}_1) d\boldsymbol{c}_1\right] \prod_{i=2}^{n} \left[\int p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i) p(\boldsymbol{c}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}) d\boldsymbol{c}_i\right] \\
&= \int \cdots \int \left[p(\boldsymbol{x}_0, \boldsymbol{c}_1, \boldsymbol{x}_1) \prod_{i=2}^{n} p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, \boldsymbol{c}_i) p(\boldsymbol{c}_i|\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2})\right] d\boldsymbol{c}_1 \cdots d\boldsymbol{c}_n \\
&= \int \cdots \int p(\boldsymbol{x}_0, \boldsymbol{c}_1, \boldsymbol{x}_1, ..., \boldsymbol{c}_n, \boldsymbol{x}_n) d\boldsymbol{c}_1 \cdots d\boldsymbol{c}_n \\
&= p(\boldsymbol{x}_0, ..., \boldsymbol{x}_n)
\end{aligned}$$

This confirms the joint distribution in Eq. (5) is consistent with the marginal AR-2 process. □

B Experimental Details

B.1 Model Checkpoints

We use the pretrained VLM models Qwen2.5-VL-3B-Instruct and InternVL2.5-8B, available at `https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct` and `https://huggingface.co/OpenGVLab/InternVL2_5-8B`, respectively. We also use the pretrained Stable Diffusion 3.0 model available at `https://huggingface.co/stabilityai/stable-diffusion-3-medium`. Evaluation is performed using the script for testing IQA (Image Quality Assessment) in `https://github.com/cswry/OSEDiff`.

B.2 User Prompts

The user prompt used for the base VLM is as follows:

> The second image is a zoom-in of the first image. Based on this knowledge, what is in the second image? Give me a set of words.

The user prompt used for the critic VLM is as follows:

> First Image: `<image>`
> Second Image: `<image>`
> The second image is a zoom-in of the first image. Please rate the quality of the following description on how well it describes the second image. Output only a single score between 0 and 100.
> Description: `<Output of Base VLM>`
> Rating (0-100):

B.3 Other Settings

The backbone SR model is trained based on the training scheme of OSEDiff [48], with Stable Diffusion 3.0 as the backbone diffusion model. We train using four NVIDIA GeForce RTX 3090 GPUs with the LSDIR [24] dataset and 10K images from FFHQ [18]. Coarse-to-fine training is used: random degradation (same setting as OSEDiff) for 25K iterations, then $4\times$ specific upscaling for 20K iterations. Other settings (*e.g.*, batch size, learning rate, etc.) follow the default settings of OSEDiff.

https://doi.org/10.52202/085713-4181

The VLM model is GRPO fine-tuned using four NVIDIA GeForce RTX 3090 GPUs with the LSDIR dataset, with a train/validation split ratio of 0.01 (*i.e.*, 849 images for validation). Specifically, the Qwen2.5-VL-3B-Instruct model is LoRA fine-tuned (Rank: 8, Alpha: 32, Dropout: 0.05), with two generations per prompt for 10K global steps. The SWIFT [59] infrastructure was used for this process. Reward graphs during training for the validation set are given in Fig. 6 of the main paper.

Evaluation is performed with the code provided in [48], modified for no-reference metric evaluation. For occasional failure cases, worst values are given for each metric (100.0 for NIQE, 0.0 for others).

C Algorithms

The following algorithms are provided:

- Algorithm 1: the main algorithm for Chain-of-Zoom inference.
- Algorithm 2: the algorithm for GRPO-based human preference alignment training of VLMs.

Algorithm 1 Chain-of-Zoom Inference

Input: Low resolution image $\boldsymbol{x}_L$, Super-resolution model p_θ, VLM p_ϕ, Number of recursions n
Output: High resolution image $\boldsymbol{x}_n$

$\boldsymbol{x}_0 \leftarrow \boldsymbol{x}_L$
for $i : 1 \rightarrow n$ **do**
  if $i = 1$ **then**
    $\boldsymbol{c}_i \leftarrow p_\phi(\boldsymbol{c}_i | \boldsymbol{x}_{i-1})$
  else
    $\boldsymbol{c}_i \leftarrow p_\phi(\boldsymbol{c}_i | \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2})$
  end if
  $\boldsymbol{x}_i \leftarrow p_\theta(\boldsymbol{x}_i | \boldsymbol{x}_{i-1}, \boldsymbol{c}_i)$
end for

Algorithm 2 GRPO-based RL Training of Prompt-Extraction VLM

Input: Base (prompt-extraction) VLM p_ϕ with parameters ϕ, Critic VLM V_{critic}, Phrase blacklist B_{phrase} for R_{phrase}, Number of training iterations N_{iter}, Number of generations per prompt N_{gen}, Training dataset $D = \{(\boldsymbol{x}_{k-2}^{(j)}, \boldsymbol{x}_{k-1}^{(j)})\}_{j=1}^{M}$ of multi-scale image crop pairs
Output: Fine-tuned prompt-extraction VLM p_ϕ

for iteration $t : 1 \rightarrow N_{\text{iter}}$ **do**
  for generation $g : 1 \rightarrow N_{\text{gen}}$ **do**
    Sample a multi-scale image pair $(\boldsymbol{x}_{i-2}, \boldsymbol{x}_{i-1})$ from D
    Generate candidate prompt $\boldsymbol{c}_i^{(g)} \sim p_\phi(\cdot | \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2})$
    $s_{\text{critic}} \leftarrow V_{\text{critic}}(\boldsymbol{c}_i^{(g)} \mid \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2})$ ▷ Critic VLM scores prompt, range $[0, 100]$
    $R_{\text{critic}} \leftarrow \text{Rescale}(s_{\text{critic}}, 0, 1)$ ▷ Rescale score to $[0, 1]$
    $R_{\text{phrase}} \leftarrow 1$
    for all $b \in B_{\text{phrase}}$ **do**
      if phrase b is in $\boldsymbol{c}_i^{(g)}$ **then**
        $R_{\text{phrase}} \leftarrow 0$
        break
      end if
    end for
    $R_{\text{rep}} \leftarrow -\text{FractionOfRepeatedNgrams}(\boldsymbol{c}_i^{(g)})$ ▷ Repetition Penalty, range $[-1, 0]$
    $R(\boldsymbol{c}_i^{(g)}) \leftarrow w_{\text{critic}} R_{\text{critic}} + w_{\text{phrase}} R_{\text{phrase}} + w_{\text{rep}} R_{\text{rep}}$ ▷ Total weighted reward
  end for
  $\hat{A}^{(g)} \leftarrow R(\boldsymbol{c}_i^{(g)}) - \frac{1}{N_{\text{gen}}} \sum_{n=1}^{N_{\text{gen}}} R(\boldsymbol{c}_i^{(n)})$ ▷ Group-based advantage estimation
  Calculate $\mathcal{L}_{\text{GRPO}}(\phi)$ with estimated advantages $\hat{A}^{(g)}$ ▷ Detailed procedure in [34]
  Update parameters ϕ of p_ϕ using GRPO policy update with $\mathcal{L}_{\text{GRPO}}(\phi)$
end for

D User Study

We further prove that GRPO fine-tuning of the VLM enhances human preference alignment by performing a MOS (mean-opinion-score) test on various samples for 25 human participants. Specifically, we compare between three different VLM prompts: (i) prompts generated from only the LR input (*i.e.*, $p_\phi(c_i \mid x_{i-1})$); (ii) prompts generated from multi-scale image prompts (*i.e.*, $p_\phi(c_i \mid x_{i-1}, x_{i-2})$); and (iii) prompts generated after GRPO fine-tuning (*i.e.*, $p_\phi(c_i \mid x_{i-1}, x_{i-2})$ with RL-trained ϕ).

Example questions are provided in Fig. 9. After being given a set of instructions, each user was asked to evaluate five different sets of randomly mixed zoom sequences and five different sets of randomly mixed text generations. Users expressed their preference from 'Very Bad' to 'Very Good', and the preferences were converted to a score of 1 to 5. Resulting preference scores are shown in Fig. 10. We further conduct pair-wise t-test to confirm the statistical significance of the scores.

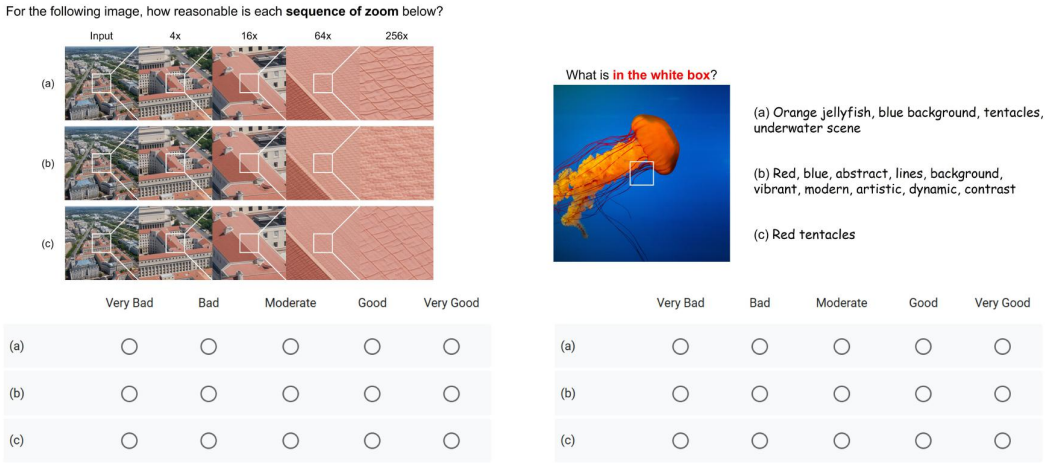


Figure 9: Example questions used for the MOS test. (Left) **Human-Preferred Image Generation.** Users were first given the instruction: 'In this survey, several samples will be given where we zoom into the center of the image. For each sequence of zoom, please rate how preferable the zoom is. (i.e., If we zoom into this input image, will the images look like this sequence?)' (Right) **Human-Preferred Text Generation.** Users were first given the instruction: 'In this section, several samples will be given where we try to explain the center of the image. For each image, please rate how preferable the explanation is. (i.e., Does the text explanation well explain what is in the white box?)'

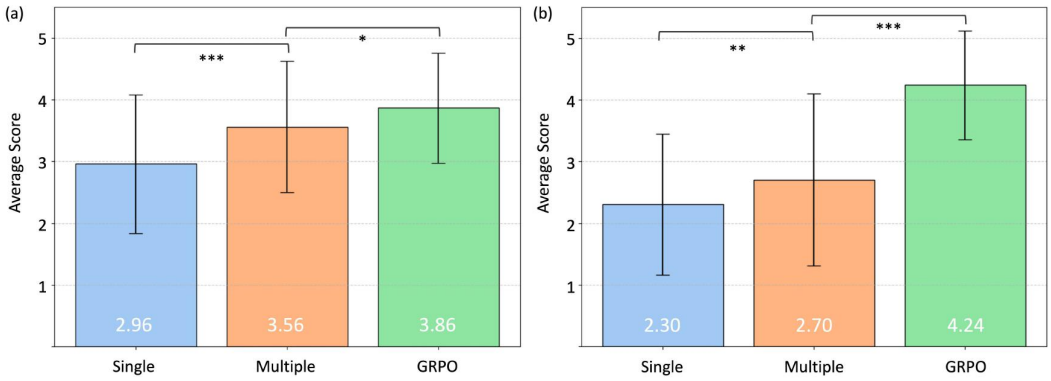


Figure 10: (a) Mean opinion scores for image generation. (b) Mean opinion scores for text generation. The scores on each bar denote the means and the error bars represent standard deviation. Significance of scores are denoted as, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

https://doi.org/10.52202/085713-4181

E Additional Results for Performing CoZ with Open-Source OSEDiff

We further prove the applicability of our CoZ framework with the open-source OSEDiff [48] model (leveraging Stable Diffusion v2.1 backbone) available at `https://github.com/cswry/OSEDiff`. Quantitative comparison on DIV2K, DIV8K training datasets are provided in Tab. 4 and example qualitative results are provided in Fig. 11. Results show that CoZ is robust and shows good performance when utilizing OSEDiff that leverages the Stable Diffusion v2.1 model as its backbone.

Table 4: Quantitative comparison using the open-source OSEDiff. **Best**, Second-Best.

		DIV2K				DIV8K			
Scale	Method	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑
4×	NN Interpolation	12.1252	39.96	0.3396	0.2630	13.1984	40.26	0.3472	0.2672
	Direct SR	4.7572	69.26	0.6366	0.7266	4.8659	68.16	0.6349	0.7198
	CoZ (Null)	4.7295	69.34	0.6359	0.7272	**4.8174**	68.11	0.6332	0.7184
	CoZ (DAPE)	4.7577	69.26	0.6366	0.7265	4.8662	68.16	**0.6350**	0.7199
	CoZ (VLM)	**4.7241**	**69.42**	**0.6368**	**0.7279**	4.8437	**68.31**	0.6346	**0.7224**
16×	NN Interpolation	22.1215	24.01	0.3378	0.2346	22.2744	24.94	0.3465	0.2585
	Direct SR	6.9951	51.88	0.5361	0.6206	7.4394	51.65	0.5472	0.6300
	CoZ (Null)	6.5369	61.86	0.5776	**0.6988**	6.7363	60.76	0.5842	0.6919
	CoZ (DAPE)	6.5628	61.47	0.5799	0.6899	6.7985	60.58	0.5888	0.6926
	CoZ (VLM)	**6.5254**	**62.05**	**0.5801**	0.6958	**6.7348**	**61.11**	**0.5904**	**0.6978**
64×	NN Interpolation	27.4051	37.69	0.3803	0.3690	27.7533	37.13	0.3861	0.3837
	Direct SR	15.6269	21.56	0.4255	0.4943	15.8252	22.02	0.4316	0.5059
	CoZ (Null)	8.9369	54.46	0.5598	**0.6672**	8.9645	53.48	0.5643	0.6655
	CoZ (DAPE)	8.8681	53.50	0.5622	0.6553	9.0221	52.76	0.5687	0.6616
	CoZ (VLM)	**8.8259**	**54.84**	**0.5645**	0.6615	**8.8553**	**53.84**	**0.5716**	**0.6677**
256×	NN Interpolation	34.8461	27.01	0.4179	0.5259	37.2612	26.98	0.4184	0.5299
	Direct SR	15.6688	26.37	0.4593	0.5203	15.9510	26.17	0.4574	0.5231
	CoZ (Null)	11.0907	47.14	0.5441	0.6223	11.0661	47.09	0.5439	0.6297
	CoZ (DAPE)	11.0014	45.81	0.5440	0.6162	10.9251	46.50	0.5475	0.6345
	CoZ (VLM)	**10.8156**	**48.22**	**0.5495**	**0.6257**	**10.7086**	**48.25**	**0.5518**	**0.6384**

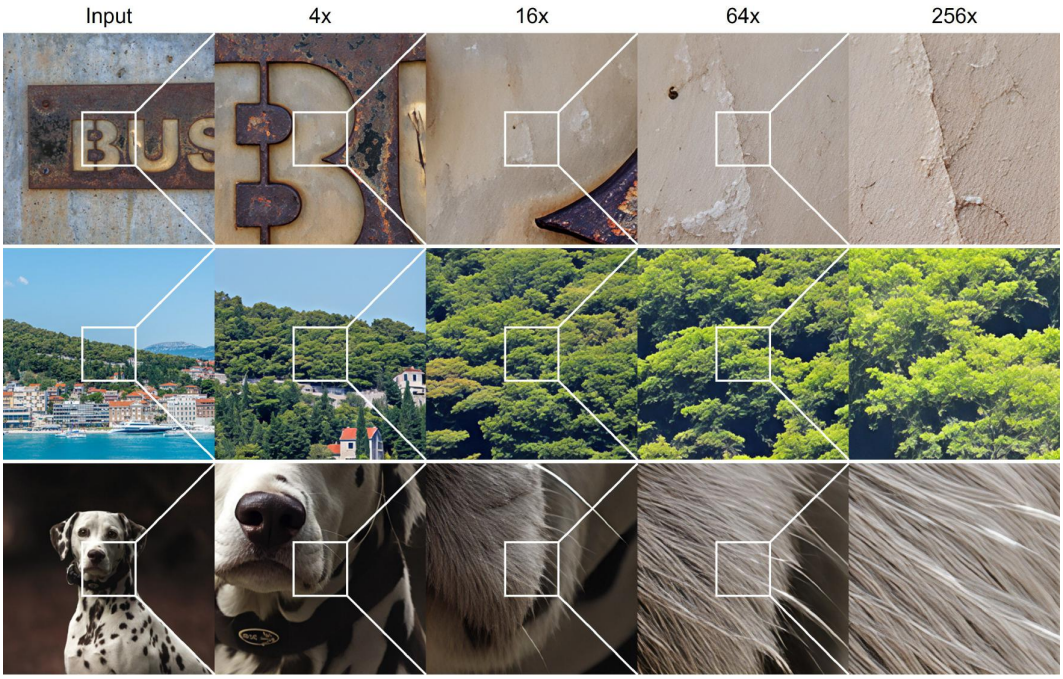


Figure 11: Qualitative results for performing CoZ with the open-source OSEDiff (leveraging Stable Diffusion v2.1 as the diffusion backbone). The GRPO fine-tuned VLM is used as the prompt extractor.

https://doi.org/10.52202/085713-4181

F Using Qwen2.5-VL-7B-Instruct as the Critic VLM

Quantitative results using Qwen2.5-VL-7B-Instruct as the critic VLM is provided in Tab. 5. All settings are identical to Tab. 1 except for choice of critic VLM.

Table 5: Quantitative results using Qwen2.5-VL-7B-Instruct as the critic VLM.

		DIV2K				DIV8K			
Scale	Method	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑
4×	Direct SR	4.7320	**67.00**	**0.6344**	**0.7005**	4.8631	**66.29**	**0.6359**	**0.6946**
	CoZ (VLM)	**4.6546**	66.89	0.6336	0.6993	**4.7947**	66.19	0.6349	0.6939
16×	Direct SR	7.2183	51.25	0.5406	0.6080	7.5855	50.17	0.5473	0.6035
	CoZ (VLM)	**6.2600**	**58.44**	**0.5955**	**0.6520**	**6.5513**	**57.62**	**0.5996**	**0.6572**
64×	Direct SR	16.5915	22.54	0.3995	0.4309	16.5874	22.97	0.4069	0.4451
	CoZ (VLM)	**7.8554**	**52.12**	**0.5824**	**0.6229**	**8.0089**	**51.07**	**0.5826**	**0.6235**
256×	Direct SR	16.1749	28.89	0.4470	0.5196	15.8667	28.90	0.4464	0.5256
	CoZ (VLM)	**8.9237**	**48.60**	**0.5786**	**0.5948**	**8.8982**	**48.06**	**0.5767**	**0.6040**

G Zooming into Overlapping Regions

Cosine similarity measurements of overlapping patches across scales are provided in Tab. 6. Specifically, for the first 500 images of the DIV2K dataset, we create overlapping patch pairs with overlapping amounts of 50% (*i.e.*, the right half of one patch and the left half of the other patch represent the same region in the original image). The pixels of overlapping patches are flattened into a 1-dimensional vector and their cosine similarity are calculated. We evaluate for patch pairs across different magnifications (multiple recursions), and observe error accumulation. Cosine similarity for non-overlapping patches are also reported as reference. Results show that cosine similarity of overlapping patches are high, even for the highest magnification of $256\times$.

Table 6: Cosine similarity for overlapping/non-overlapping regions.

	Overlapping Regions			Non-overlapping Regions		
Scale	Mean	Median	Minimum	Mean	Median	Minimum
4×	0.9977	0.9982	0.9864	0.7951	0.8196	0.3077
16×	5.5208	6.6210	0.5387	0.5535	5.5940	0.5346
64×	4.9803	6.6210	0.6361	0.6596	5.0305	0.6328
256×	4.8637	6.6210	0.6459	0.6835	4.9414	0.6405

H Full Image Super-Resolution

Chain-of-Zoom can be applied to super-resolution of full images by using the same tiling method used in prior works (*e.g.*, SeeSR, OSEDiff). Specifically, we use tiling of VAE and tiling of latents as described below to produce full-image super-resolution results without boundary artifacts.

1. A given LR image is resized to the target resolution and VAE encoding is done in tiles, allowing encoding to be performed even in settings of limited GPU memory.
2. The encoded (low-resolution) latent is tiled into overlapping patches. For latent sizes of 64×64, we find overlaps of 16 to work sufficiently well.
3. Each low-resolution patch of 64×64 passes through the super-resolution network to become high-resolution patches, each guided by patch-specific prompts generated by the prompt-extractor VLM. Note that this step requires multiple passes of the VLM, a computational bottleneck to be solved by future work.
4. The output high-resolution patches are multiplied by Gaussian weights in overlapping regions for smooth transposition between patches, and then combined to create the final high-resolution image.
5. The whole process is repeated as *scale autoregression* to achieve higher resolutions.

https://doi.org/10.52202/085713-4181

I Additional Qualitative Results

Additional qualitative results of extreme super-resolution by CoZ are provided below.

Figure 12: Extreme super-resolution of photorealistic images by CoZ up to $64\times$ magnification.

https://doi.org/10.52202/085713-4181

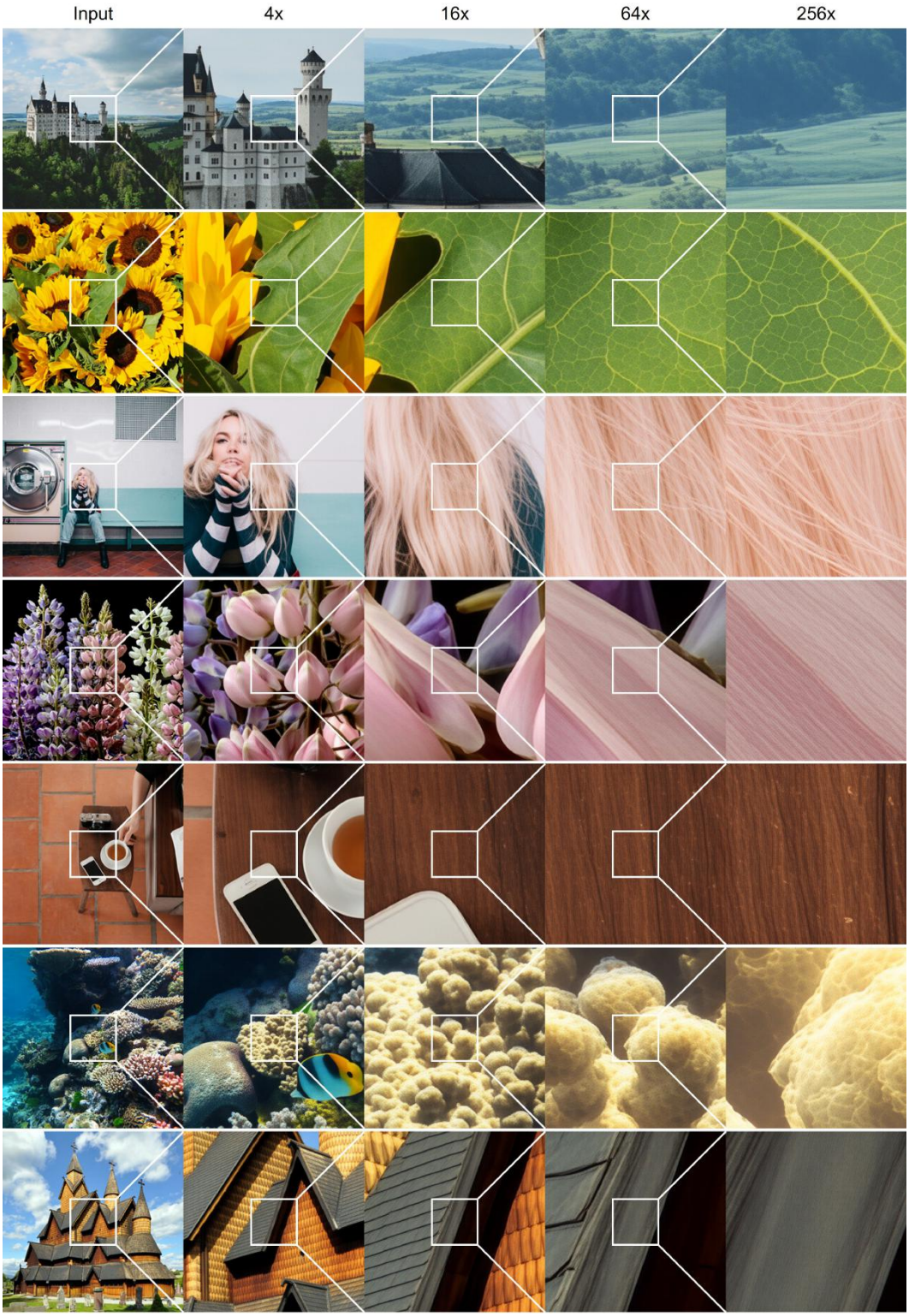


Figure 13: Extreme super-resolution of photorealistic images by CoZ up to $256\times$ magnification.

https://doi.org/10.52202/085713-4181

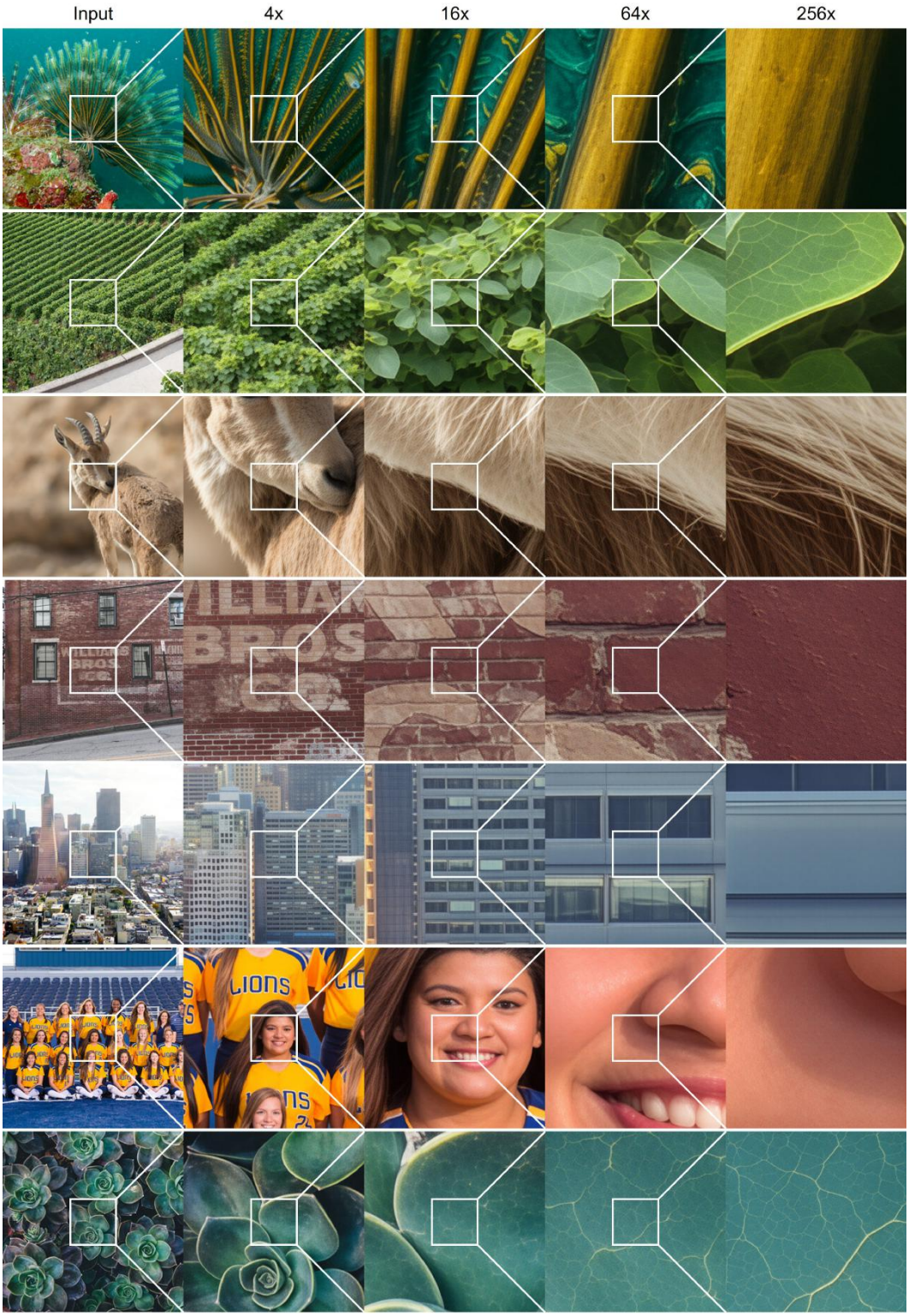


Figure 14: Extreme super-resolution of photorealistic images by CoZ up to $256\times$ magnification.

https://doi.org/10.52202/085713-4181

Figure 15: Extreme super-resolution of photorealistic images by CoZ up to $256\times$ magnification.

https://doi.org/10.52202/085713-4181

Figure 16: Extreme super-resolution of photorealistic images by CoZ up to $256\times$ magnification.

https://doi.org/10.52202/085713-4181

J Example Failure Modes of VLM before Fine-Tuning

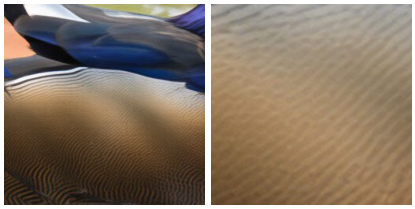
J.1 Repetition

The second image is a zoom-in of the first image. Based on this knowledge, what is in the second image? Give me a set of words.

RESPONSES

Qwen2.5-VL-3B-Instruct	fur texture orange background animal fur close-up pattern texture orange fur texture orange fur background orange fur texture orange fur background orange fur texture orange fur texture orange fur texture orange fur texture orange fur texture orange fur texture orange fur texture orange fur texture orange fur texture orange fur texture orange fur texture orange fur ...
+ GRPO Training	fur

J.2 Unwanted Phrases



The second image is a zoom-in of the first image. Based on this knowledge, what is in the second image? Give me a set of words.

RESPONSES

Qwen2.5-VL-3B-Instruct	The second image shows a close-up view of a surface with a textured pattern. The texture appears to be a combination of smooth and slightly raised areas, giving it a somewhat wavy or ripple-like appearance. The color gradient ranges from a lighter shade at the top to a darker shade at the bottom, creating a sense of depth and dimension.
+ GRPO Training	feathers

https://doi.org/10.52202/085713-4181

J.3 Suboptimal Results

The second image is a zoom-in of the first image. Based on this knowledge, what is in the second image? Give me a set of words.

RESPONSES

Qwen2.5-VL-3B-Instruct	ant leg
+ GRPO Training	crab claw

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions are a scale-level autoregressive framework and a novel RL pipeline, which are both explained in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have created a separated Limitations section to discuss the limitations of our work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/085713-4181

Answer: [Yes]

Justification: For each of our theoretical results, we provide the assumptions we make and provide proof, also included in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide accurate information of datasets and models used in our experiments. Experimental settings are provided both in the main paper and in supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use datasets that are open to access, and specific codes are provided. We further provide sufficient information for reproduction in the supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental settings are included in the core of the paper, and further details are provided in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/085713-4181

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information about the type of compute workers used for the experiments in supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not violate the NeurIPS Code of Ethics in any way.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the Limitations section, we also discuss the potential societal impacts of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any pretrained language models, image generators, or scraped datasets that have a high risk for misuse; we only introduce a framework for utilizing pretrained models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide citations for the code, data, models used, and properly respect the licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

 https://doi.org/10.52202/085713-4181

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not provide any specific assets, but rather a framework for utilizing pretrained models.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide the full text of instructions given to participants for user study in the supplemental material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not perform any studies that require IRB approvals.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

 https://doi.org/10.52202/085713-4181