Data-efficient Large Vision Models through Sequential Autoregression

Zhiwei Hao [* 1] **Jianyuan Guo** [* 2] **Chengcheng Wang** [* 3]
Yehui Tang [3] **Han Wu** [2] **Han Hu** [✉ 1] **Kai Han** [3] **Chang Xu** [✉ 2]

haozhw@bit.edu.cn; {jianyuan.guo,han.wu}@sydney.edu.au; {chengcheng.wang,yehui.tang,kai.han}@huawei.com

Abstract

Training general-purpose vision models on purely sequential visual data, eschewing linguistic inputs, has heralded a new frontier in visual understanding. These models are intended to not only comprehend but also seamlessly transit to out-of-domain tasks. However, current endeavors are hamstrung by an over-reliance on colossal models, exemplified by models with upwards of 3B parameters, and the necessity for an extensive corpus of visual data, often comprising a staggering 400B tokens (Bai et al., 2023). In this paper, we delve into the development of an efficient, autoregression-based vision model, innovatively architected to operate on a limited dataset. We meticulously demonstrate how this model achieves proficiency in a spectrum of visual tasks spanning both high-level and low-level semantic understanding during the testing phase. Our empirical evaluations underscore the model's agility in adapting to various tasks, heralding a significant reduction in the parameter footprint, and a marked decrease in training data requirements, thereby paving the way for more sustainable and accessible advancements in the field of generalist vision models. The code is available at https://github.com/ggjy/DeLVM.

1. Introduction

Training a generalist model capable of executing diverse tasks simultaneously—and with the agility to tackle new tasks given just few examples—represents a pivotal stride toward artificial general intelligence within the computer vision community. In the realm of contemporary natural language processing (NLP), large language models trained via autoregression, such as GPT (Brown et al., 2020; Achiam et al., 2023), have demonstrated the remarkable ability to comprehend and generate natural language text, particularly excelling in complex and nuanced contexts. These models leverage language sequences as a universal interface, facilitating rapid adaptation to a variety of language-centered tasks with minimal prompting and examples.

However, the landscape of computer vision is markedly different. Unlike the uniformity of input-output structures in language tasks, visual tasks exhibit a rich diversity in their formats, posing significant challenges to the development of comparable generalist models. Recently, advances in LVM (Bai et al., 2023) have redefined 'visual sentences', enabling the representation of both raw images and annotated data without requiring meta-knowledge beyond the pixel level. This paradigm shift opens a new vista for the evolution of generalist models within the visual domain.

Nevertheless, while the capability of LVM is bolstered by the employment of large-scale datasets, such as more than 1.6 billion images, this reliance simultaneously engenders complexities in the training process of LVM. One obstacle arises from the imbalance of datasets across different tasks. Our visual world naturally exhibits a long-tailed distribution of different tasks. The model performance on tasks with limited data representation is substantially impaired when there is an uneven distribution of data across tasks, with certain tasks having an abundance of data and others suffering from a deficiency. For example, the segmentation benchmark SA-1B (Kirillov et al., 2023), containing 11 million images, provides a wealth of data compared to the keypoint detection benchmark COCO (Lin et al., 2014), containing only 0.2 million images. Direct training on the combined dataset of these two benchmarks results in the model's inability to learn keypoint detection due to the overwhelming amount of segmentation data, as iilustrated in Figure 1. This observation forms the basis of our straightforward data augmentation strategy, which automatically

---

[*]Equal contribution [1]School of information and Electronics, Beijing Institute of Technology, Beijing, China [2]School of Computer Science, Faculty of Engineering, University of Sydney, Sydney, Australia [3]Huawei Noah's Ark Lab, Beijing, China. Correspondence to: Han Hu <hhu@bit.edu.cn>, Chang Xu <c.xu@sydney.edu.au>.
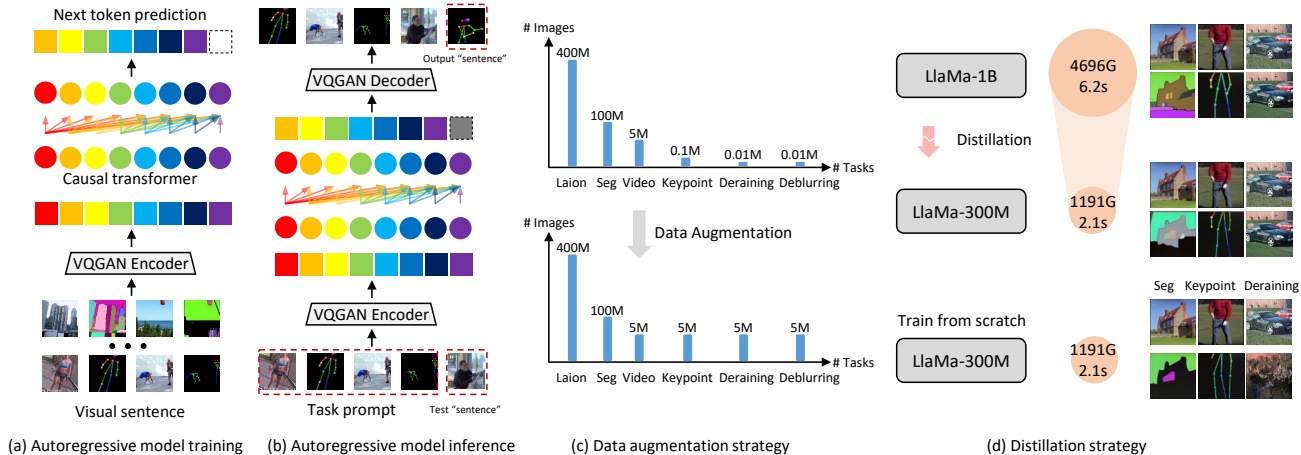
Figure 1: An overview of our framework. We follow the autoregressive setting proposed in LVM (Bai et al., 2023), which encodes input images into a 1D sequence. We further explore the data-efficient LVMs via data augmentation and knowledge distillation strategy.

enriches smaller datasets and achieves balance among diverse tasks. Specifically, we enhance underrepresented (tail) datasets by randomly augmenting the training samples. We find that this approach yields stronger performance compared to traditional re-sampling strategy.

The significant performance gain brought by the proposed data augmentation strategy compels us to contemplate: Given that a substantial reduction in the volume of data required for training does not adversely affect model efficacy, could there potentially exist superfluousness within the model's parameterization? We resort to knowledge distillation (KD) methods to enhance the performance of compact LVMs. In our study, we demonstrate the effectiveness of KD in autoregressive LVMs. While models with larger capacity initially exhibit superior performance and surpasses the student model trained from scratch by a substantial margin, the incorporation of KD significantly narrows the performance gap and shed light on how to build efficient LVMs.

In general, this paper explores the development of data-efficient autoregressive-based large vision models (DeLVMs). Our study focuses on data augmentation strategy, especially in scenarios with long-tail distributions across different visual tasks. We demonstrate that simple data augmentation techniques yield satisfactory results compared to re-sampling baselines. We also leverage KD to create more compact and efficient LVMs, which leads to a significant reduction in validation loss, with improvement in accuracy and decrease in perplexity. This highlights the potential of KD to bridge the performance gap and enhance the capabilities of autoregressive LVMs. Our work aims to advance the understanding and construction of more efficient autoregressive vision models capable of simultaneously addressing various vision tasks.

2. Autoregressive Large Vision Models

Inspired by the achievements of LLMs (Brown et al., 2020; Touvron et al., 2023b; Chowdhery et al., 2023), various endeavors have been made to develop autoregressive models specifically tailored for vision tasks. An outstanding breakthrough in this domain, referred to as the LVM (Bai et al., 2023), transforms visual data into *visual sentences*, thereby enabling the uniform modeling of diverse vision tasks. A typical LVM consists of two key modules: (1) a VQGAN (Esser et al., 2021) for input image tokenization; (2) a transformer model (Touvron et al., 2023a) for sequential autoregressive prediction.

VQGAN tokenizer. The initial step of modeling images via a transformer-based autoregressive approach involves converting original images into discrete tokens. A VQGAN (Esser et al., 2021) is utilized for this tokenization process, composed of three main components: an encoder, a decoder, and a trainable codebook. When training a VQGAN model, the encoder takes an image as input and projects it into a grid of features, which are then mapped to specific codes in the codebook. The decoder is used to reconstruct the original image based on the grid of codes. A pretrained VQGAN model is capable of converting the input image into a sequence of tokens following the scan-line order, which can be modeled by an autoregressive model.

Autoregressive model. After tokenization, visual sentences are constructed by concatenating tokens from multiple images and fed into a causal transformer model which is employed for autoregressive prediction. The goal of the causal transformer is to predict the next token at each position, using cross-entropy as the loss function. Specifically, considering an input sentence of length L as $s_i = \{x_1, x_2, ..., x_{L-1}, x_L\}$, where x_l represents the l-th token, the objective of the causal transformer is to predict
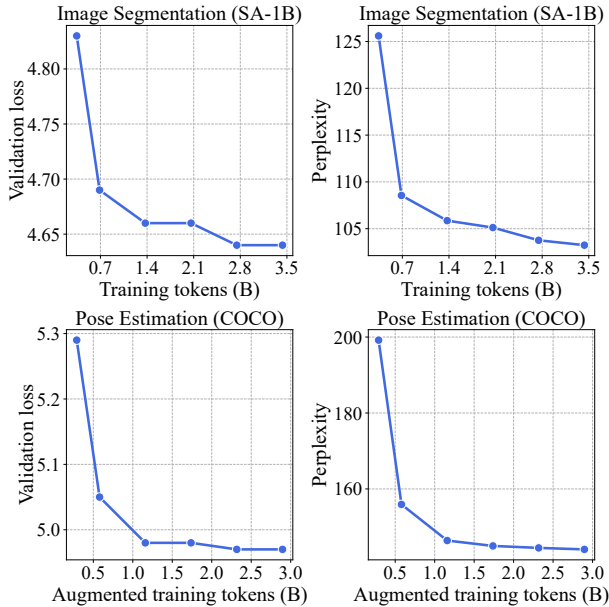
Figure 2: **Data augmentation yields a similar effect to the introduction of new data.** We train a LLaMA-300M model on subsets of SA-1B and augmented COCO-Pose datasets with a fixed 12K training steps. In segmentation task with ample available data, enhanced model performance is observed upon introducing more training data. Conversely, for human pose estimation task with limited data, augmenting the dataset has a comparable impact to the introduction of new training data. The original COCO-Pose dataset contains approximately 0.03 billion tokens.

$$s_o = \{x_2, x_3, ..., x_L, \varnothing\}.$$

Prompted inference. After training, LVM is capable of performing inference on downstream tasks via vision prompting. Firstly, several input-output image pairs are converted into a visual sentence, which serves as a task definition. This sentence is then combined with the test image, which is also tokenized via the VQGAN encoder. Based on the concatenated sentence, the model generates output tokens based on the input sequential tokens. The final predicted image is then obtained by decoding above generated tokens using the decoder of the VQGAN model.

3. Data Augmentation

In common practices of training highly proficient LLMs, the training duration is often restricted to just one epoch to mitigate the risk of overfitting (Brown et al., 2020; Touvron et al., 2023a;b). In contrast to the NLP field, where numerous large-scale corpora are readily available, many computer vision tasks encounter a scarcity of extensive datasets. The scarcity makes it impractical to directly transfer the training

schedule settings from language tasks to vision tasks. Fortunately, various data augmentation techniques commonly employed in traditional computer vision tasks can be leveraged to enhance the training data for LVMs. We initiate the examination of data augmentation effectiveness by training LVM on various vision tasks.

3.1. Single Task

We initiate our study by conducting experiments on a single task. It is intuitive that the introduction of new training data enhances overall model performance. To delve deeper into this effect, we proceed to train the model on an image segmentation task using varying amounts of available data. Following the configuration of LVM (Bai et al., 2023), we employ a causal LLaMA model (Touvron et al., 2023a) with 300M parameters for autoregressive modeling. For tokenization, we utilize an off-the-shelf VQGAN trained by Chang *et al.* (Chang et al., 2023), which translates each image into 256 discrete tokens. The VQGAN has a codebook size of 8192, with each code in the codebook having a dimension of 64. To align the dimension of the code with that of the transformer model, we insert a learnable linear layer between these two modules. Regarding the dataset, we utilize various subsets of the SA-1B (Kirillov et al., 2023), spanning from 1% to 10%. The training steps for all models remain fixed at 12K. Cross-entropy loss and perplexity on a withheld subset of SA-1B (equivalent to 1% of the dataset) serve as the metrics. The top two figures in Figure 2 illustrate the impact of available data on model performance. Obviously, as more new data is introduced during training, there is a significant increase in model performance. For instance, when the utilized data from SA-1B increases from 1% to 10% (0.34B tokens to 3.43B tokens), the validation loss decreases by 0.19, and perplexity decreases by 22.4.

However, in certain tasks, such as human pose estimation, the availability of data is constrained, making it challenging to enhance model performance by introducing new samples. To bridge this gap, we employ data augmentation techniques to augment the existing dataset. Specifically, for each sample in the dataset, we apply random crop and random flip operations to generate several augmented versions. These augmented samples are then tokenized by the VQGAN model and incorporated into training alongside the tokenized original dataset. We assess the impact of augmentation by varying the augmentation range from 10 to 100 times. As depicted in the bottom two figures in Figure 2, expanding the data scale through data augmentation also contributes to improved model performance. The decreasing trend observed in validation loss and perplexity as the augmentation times increase mirrors the pattern observed when introducing new samples. This outcome underscores that data augmentation has a comparable effect to acquiring more new data, offering a straightforward approach to

effectively train LVMs in data-limited scenarios.

3.2. Multiple Tasks

In practical scenarios, the goal is to train a versatile LVM capable of handling multiple tasks. However, data imbalance poses a challenge, with some tasks having ample data while others have limited data. Training directly with this unbalanced mixture of data from various tasks hinders overall model performance. To address this issue, the re-sampling scheme (Zhang et al., 2023) repeat samples from minority classes to rebalance the classes. Following this approach, we balance the data amount of different tasks by repeating samples from tasks with limited data. Additionally, we explore the effectiveness of using data augmentation to achieve balanced task data. To conduct multitask experiments, we focus on three distinct vision tasks: image segmentation, human pose estimation, and image deraining. The training data for these tasks comprises a subset of the SA-1B dataset (approximately 10% of the entire dataset) (Kirillov et al., 2023), the complete COCO-Pose dataset (Lin et al., 2014), and the entire Rain13K dataset (Jiang et al., 2020). In each data setting, the model undergoes a fixed total of 35K iterations during the training process. To evaluate the performance of our trained model, we use a withheld subset of SA-1B, along with the MPII dataset (Andriluka et al., 2014) and the Test2800 dataset (Fu et al., 2017b).

Table 1 lists the validation results. In comparison to training on the direct mixture of data from the three tasks, training the model on data balanced by re-sampling even results in worse performance, particularly evident in the human pose estimation and image deraining tasks with repeated samples. By contrast, the model trained on the augmented dataset achieves better performance, notably excelling in the human pose estimation and image deraining tasks. Notably, the use of unbalanced data yields the best quantitative results on the image segmentation task, likely because the majority of the training data belongs to this specific task.

To closely examine the impact of various dataset processing schemes in the multitask scenario, we visualize the inference results of the trained models in Figure 3. When trained with unbalanced data, the model excels in the image segmentation task. However, for tasks of human pose estimation and image deraining, where there is less training data, its performance is subpar. For instance, in test case 2 of the deraining task, the model seems to be confused by the abundance of segmentation samples in the training data, resulting in a segmentation output despite the presence of deraining prompts. This emphasizes that the direct use of unbalanced data for training cannot yield excellent LVMs. When the model is trained with task data balanced by re-sampling, it achieves even poorer performance. In the image segmentation and deraining tasks, the model can only pro-

duce tolerable inference results in test case 1 while failing to provide informative outputs in test case 2. Moreover, in the deraining task in both cases, it yields disordered results. Conversely, when employing data augmentation to achieve data balance, the trained model demonstrates proficiency in all three tasks and outperforms the other two models.

4. Knowledge Distillation

In the quest for efficient and compact models, knowledge distillation (KD) stands out as a prevalent technique used to bolster model performance, as highlighted in (Hinton et al., 2015). KD capitalizes on a pre-trained, larger teacher model to guide a smaller, more efficient student model to emulate the teacher's outputs (Hao et al., 2023). Yet, the majority of KD applications have been predominantly associated with CNN and Transformer, leaving a noticeable void in the exploration of KD for autoregerssive LVMs. Addressing this oversight, our work delves into the feasibility of applying KD in the training of LVMs, aiming to extend the benefits of KD to these streamlined models.

We begin our investigation by examining the impact of KD in single-task settings. For the teacher models, we utilize LLaMA-1B (Touvron et al., 2023a), training them on specific tasks such as image segmentation and human pose estimation. Subsequently, we train LLaMA-300M as the student models, adhering to the KD framework initially proposed by Hinton et al. (Hinton et al., 2015). All additional experimental setting are maintained as described in the preceding section, ensuring consistency in our study. The results of single-task KD are summarized in Table 2. LLaMA-1B teacher models, boasting the highest number of model parameters, demonstrates superior performance on both tasks, outperforming LLaMA-300M by a significant margin. Nevertheless, with the incorporation of KD, the performance gap markedly diminishes in terms of validation loss, accuracy, and perplexity. This underscores the efficacy of KD in the single-task scenario.

Correspondingly, the inference outcomes of all the models trained for both tasks are illustrated in Figure 4. The visual results align with the quantitative data, demonstrating that the LLaMA-1B models, serving as the teacher models, present superior visualizations with a higher level of detail in the output images. On the other hand, the LLaMA-300M models, when trained from scratch (without the guidance of KD), yield inference results that are less detailed, and these are evidently surpassed by those models that have been trained utilizing the KD approach.

We expand our investigation to encompass multi-task scenarios. By adhering to the experimental settings previously outlined, we employ a LLaMA-1B teacher to impart its knowledge to a LLaMA-300M student. This knowledge

Table 1: **Balancing datasets across multiple tasks through augmentation enhances performance.** We train a LLaMA-300M on a mixed dataset involving three tasks. Three configurations for handling the unbalanced dataset are compared. Compared to directly training on the unbalanced dataset, mitigating the long-tail distribution through re-sampling led to inferior results. Conversely, achieving balance through dataset augmentation yielded the best overall performance. "-" indicates collapsed results (larger than 10^8).

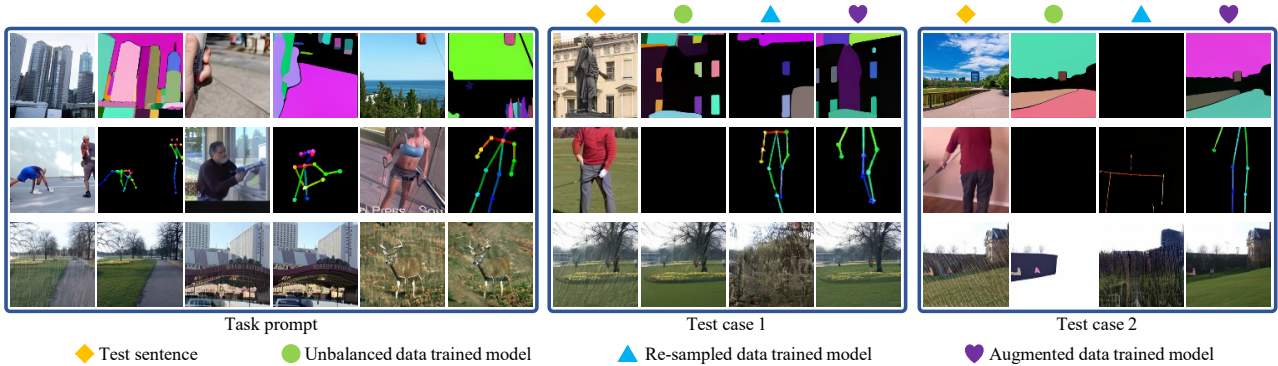| Dataset configuration	Image Segmentation (10% of SA-1B: 3.434B tokens)			Pose Estimation (COCO-Pose: 0.029B tokens)			Image Deraining (Rain13K: 0.007B tokens)		
	loss ↓	accuracy ↑	perplexity ↓	loss ↓	accuracy ↑	perplexity ↓	loss ↓	accuracy ↑	perplexity ↓
Unbalanced	4.48	20.32	88.59	4.95	19.86	141.91	5.69	11.89	279.88
Balanced by re-sampling	5.01	16.68	151.66	16.11	15.75	-	21.73	4.53	-
Balanced by augmentation	4.68	18.79	107.77	4.94	20.24	140.39	5.64	11.57	269.04

Figure 3: Visualization of inference results generated by models trained on datasets with different balancing schemes. The model trained on an unbalanced dataset exhibits biased performance, excelling primarily in the image segmentation task. Rectifying this data imbalance through re-sampling have unfortunately led to a decrease in performance. Conversely, employing augmentation to balance the dataset results in improved visualization outcomes.
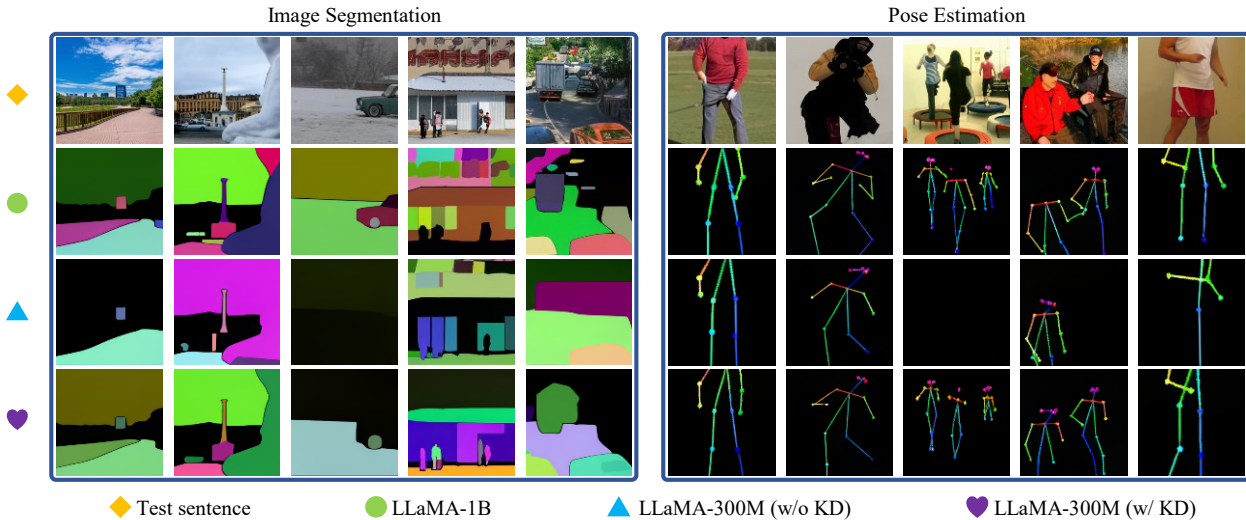

Figure 4: Visualization of the inference results of single-task trained models. The LLaMA-1B teacher model produces the best outcomes, and the LLaMA-300M model trained with KD exhibits greater similarity to the teacher model compared to the model trained from scratch.

transfer occurs through the utilization of a combination of three datasets, all balanced via data augmentation techniques. The validation outcomes for this multi-task scenario are summarized in Table 3. Upon comparing the perfor-

mance of the distilled model with that of the model trained from scratch, we observe that the former exhibits superior results. This observation highlights the efficacy of KD in enhancing the performance of LVMs, even within the more

Table 2: **KD proves beneficial in enhancing the single-task performance of LVMs.** We employ a LLaMA-1B model as the teacher to train a student model LLaMA-300M using KD. In comparison to training the student model from scratch, the introduction of KD significantly improves performance on both the image segmentation and human pose estimation tasks.

Model	KD	loss ↓	accuracy ↑	perplexity ↓
Image Segmentation (10% of SA-1B)				
LLaMA-1B	-	4.50	20.18	90.04
LLaMA-300M	✗	4.64	19.17	103.24
LLaMA-300M	✓	4.59	19.48	98.72
Pose Estimation (COCO-Pose)				
LLaMA-1B	-	4.90	20.96	134.07
LLaMA-300M	✗	4.97	20.46	144.08
LLaMA-300M	✓	4.91	20.80	135.91

practical and complex context of multi-task scenarios.

In addition, we conduct a quantitative evaluation of our distilled models on a foreground segmentation task. The objective is to binary-segment a given query image into foreground and background components. The task prompt consists of three example image pairs, followed by a test (query) image. We prompt our model to generate the next 256 tokens, which are then decoded into the output image. For benchmarking, we utilize the Pascal-5i dataset established in (Shaban et al., 2017), following the methodologies of (Bai et al., 2023; Bar et al., 2022). This dataset encompasses four distinct image splits, each containing between 346 and 725 images alongside corresponding segmentation masks. Each class in the dataset is represented by several image-mask pairs and is supplemented with held-out image queries for evaluation. We report our results using the mean Intersection Over Union (mIOU) metric. We examine the performance of our models in both a zero-shot setting and after finetuning. As illustrated in Table 4, the distilled LLaMA-300M model outperforms its counterparts that were trained from scratch. Despite our model being trained on the SAM dataset, where object colors are randomized, and our use of prompt images with black backgrounds, our generated images still exhibit some randomness in color. This randomness make it challenging to separate the binary segments through post-processing. To address this, we fine-tuned the model using the training split of Pascal-5i, which features uniformly black backgrounds. After finetuning, we observed a significant boost in the mIOU scores.

It is important to highlight that both LVM (Bai et al., 2023) and Visual Prompting (Bar et al., 2022) have retrained their VQGAN encoders on custom datasets, resulting in more satisfactory mIoU scores. Attaining a higher mIoU is not the primary goal of this paper; hence, we have chosen to

use the encoder trained on the Laion dataset (Schuhmann et al., 2022) directly. We believe that a better encoder and decoder would yield improved results.

5. Ablation Study

Influence of Different Prompts. In this section, we investigate the influence of different input prompts on the generated outputs in the context of foreground segmentation task. Our training dataset, SA-1B, features ground truth where each object is composed of random colors. As shown in Figure 5(a), the segmentation ground truth for example image pairs includes a pink background. In these cases, the model tends to generate outputs with a similar pink background. Conversely, if we use a prompt with a black background, as illustrated in Figure 5(b), the resulting image is likely to have a black background. A simple post-processing step can be applied to the output image by converting it to a grayscale image and then applying a very low threshold to produce a binary mask.

Furthermore, we finetune[1] our model using the training split of foreground segmentation benchmarks to focus more on the primary objects while disregarding background elements such as grass and sky. As evident from Figure 5(d), the output of the model after fine-tuning is more concise and refined. This demonstrates the potent capability of autoregressive vision models to transfer learning effectively through a pretrain-then-finetune strategy.

Continual learning. In our preceding experiments, LVMs are trained with shuffled data to address multiple tasks. To investigate the impact of the data shuffle operation, we train a LLaMA-300M model using an ordered concatenation of multi-task data. Specifically, the training data follow the sequence of SA-1B, COCO-Pose, and Rain13K. Upon completing the training for each task, we evaluate the perplexity of the model across all three tasks and present the results in Table 5. The results reveal that the model attains low perplexity exclusively on the most recently encountered task, exhibiting subpar performance on the other tasks, regardless of whether it has undergone training for those specific tasks.

We present visualizations of model inference results in Figure 6, employing the same task prompts as in previous sections. Notably, the model tends to prioritize the inference of its last-trained task for the given inputs, seemingly disregarding the guidance provided by the task prompts. These outcomes underscore the significance of the shuffle operation in the multi-task training of LVMs. Essentially, LVMs face challenges related to catastrophic forgetting in

---

[1]Given the limited number of training images-only a few hundred—we resampled these images 100 times to construct sequential data akin to what was used during training. The entire finetuning process was completed in approximately 5 minutes.

Table 3: **KD also helps enhance the multi-task performance of LVMs.** We assess the effectiveness of KD with balanced data for three tasks using augmentation. KD demonstrates its usefulness by improving the performance of the student model across all tasks.

Model	KD	Image Segmentation			Pose Estimation			Image Deraining		
		loss ↓	accuracy ↑	perplexity ↓	loss ↓	accuracy ↑	perplexity ↓	loss ↓	accuracy ↑	perplexity ↓
LLaMA-1B	-	4.55	19.72	94.75	4.86	20.77	129.95	5.53	12.20	245.33
LLaMA-300M	✗	4.68	18.79	107.76	4.94	20.24	140.39	5.63	11.57	271.38
LLaMA-300M	✓	4.67	18.84	106.81	4.93	20.32	139.27	5.62	11.93	269.04

Table 4: Results on Foreground Segmentation. † indicates we finetune the autogressive model to generate black background.

Model	Foreground Segmentation ↑			
	Split 0	Split 1	Split 2	Split 3
BEiT (IN-21k) (Bao et al., 2021)	0.38	0.93	0.90	0.95
VQGAN (IN-1k) (Esser et al., 2021)	6.96	10.55	9.59	9.43
MAE-VQGAN (IN-1k) (Bar et al., 2022)	2.22	7.07	5.48	6.28
BEiT (Figures) (Bar et al., 2022)	5.38	3.94	3.20	3.29
VQGAN (Figures) (Bar et al., 2022)	12.56	17.51	14.27	15.06
MAE-VQGAN (Figures) (Bar et al., 2022)	27.83	30.44	26.15	24.25
LVM-3B (Figures) (Bai et al., 2023)	48.94	51.29	47.66	50.82
LLaMA-300M (Laion)	12.46	16.92	12.99	15.76
Distilled LLaMA-300M (Laion)	14.72	17.91	14.55	17.13
Distilled LLaMA-300M (Laion)†	18.58	21.32	19.90	21.08

continual learning scenarios.

Table 5: **LVM suffers from catastrophic forgetting in continual learning scenarios.** We train a LLaMA-300M without shuffling training data. The model exhibited proficient result exclusively on the task corresponding to the most recently used training data.

Data order	Validation perplexity ↓		
	Segmentation	Pose Estimation	Deraining
SA-1B	102.06	243.09	1165.21
COCO-Pose	416.21	134.12	2633.08
Rain13K	1449.42	1285.80	287.79

6. Practical LLaMA-80M

To delve deeper into the performance capabilities of efficient LVMs, we augment our exploration by training LLaMA-80M models using both data augmentation and KD. To procure a powerful LLaMA-1B teacher, additional unlabeled image and video data are integrated into the training process. Table 6 lists the validation perplexity results of the trained models, revealing that the LLaMA-80M model trained with KD surpasses its counterpart trained from scratch.

We also evaluate the image understanding abilities of LLaMA-80M on ImageNet. Following the self-supervised MAE framework (He et al., 2022), we replace the VQGAN encoder with a patch embedding layer and incorporate an

(a) LLaMA-300M trained from scratch.

(b) LLaMA-300M trained from scratch.

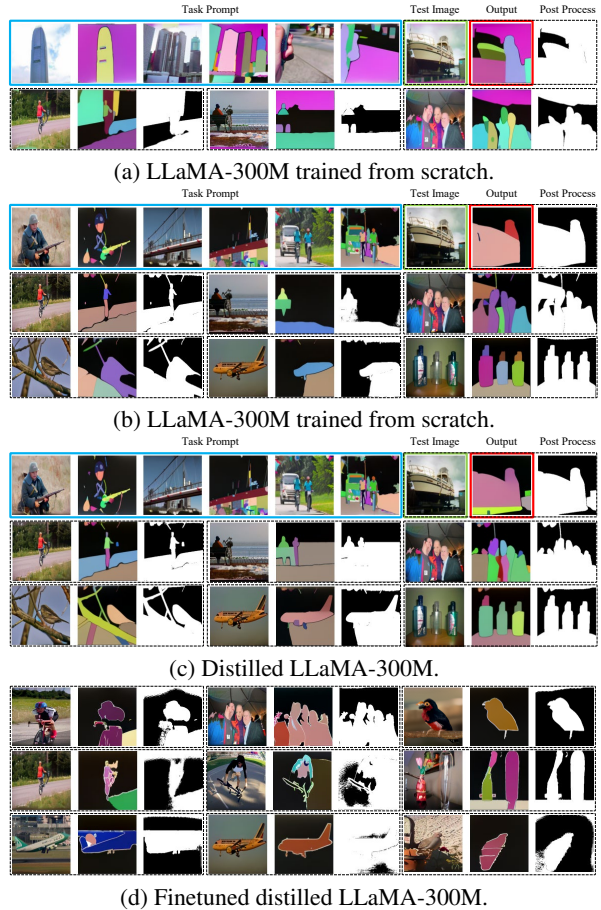(c) Distilled LLaMA-300M.

(d) Finetuned distilled LLaMA-300M.

Figure 5: Generated output and the corresponding foreground segmentation results after our post processing. Task prompt contains a sequence of images interleaved with annotations, followed by a test image. Prompts in (a) are with a pink background, in (b) (c) (d) are with a black background.

average pooling layer followed by a fully connected layer to perform the image classification task. Surprisingly, our LLaMA-80M achieves an impressive top-1 accuracy of 83% on ImageNet. Although this accuracy is lower compared to some Masked Image Modeling-based methods, it outperforms training from scratch approaches. This suggests that there may be a potential for simultaneous learning of both generation and understanding tasks, indicating a promising
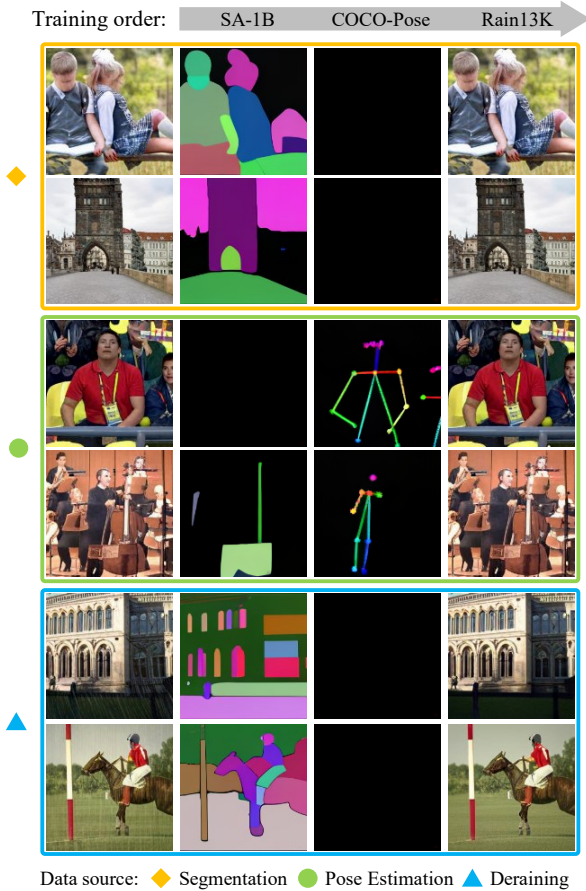
Figure 6: Visualization of inference results at different training stages. Task prompt is the same as in previous visualizations. The model's proficiency is focused solely on the most recently encountered training task, regardless of the provided prompt.

Table 6: Comparison of validation perplexity between LLaMA-80M models trained with and without KD.

Model	KD	Validation perplexity ↓		
		Segmentation	Pose Estimation	Deraining
LLaMA-1B	-	88.33	111.95	230.94
LLaMA-80M	✗	156.19	164.57	264.56
LLaMA-80M	✓	147.78	159.78	256.65
LLaMA-80M	✓	ImageNet Top-1 Acc: 83.04		

relationship between these two aspects.

7. Related Works

Autoregressive Models. Autoregressive (AR) models are central to sequence prediction and have undergone significant evolution. In the realm of natural language processing (NLP), LSTM networks emerged as foundational AR models, adept at managing long-range dependencies crucial for tasks such as language modeling. The advent of

the transformer architecture (Vaswani et al., 2017; Wang et al., 2023a; Guo et al., 2022) marked a pivotal change in AR modeling. Its self-attention mechanisms facilitated a more efficient and parallel processing approach, which in turn, catalyzed the advancement of generative pre-training methodologies (Dai & Le, 2015; Brown et al., 2020) that predict segments of text based on preceding ones. Capitalizing on their triumph in NLP, AR models have been tailored for generative image modeling tasks (Uria et al., 2013; Van Den Oord et al., 2016; Parmar et al., 2018). And many transformer-based AR vision models (Chen et al., 2020; Bai et al., 2023; Bar et al., 2022; Yu et al., 2021) have achieved notable comprehension outcomes, thus validating the efficacy of AR principles for vision-related applications.

Despite the trend, much of the existing literature primarily focuses on the advantages of up-scaling model (El-Nouby et al., 2024; Kolesnikov et al., 2019) and dataset sizes (He et al., 2022; Bai et al., 2023). While larger models exhibit superior performance, their heavy architectures demand extensive computational resources, thereby limiting their deployment on computation-constrained edge devices. Addressing this gap, our paper proposes a strategy aimed at crafting a more compact AR model that leverages data-efficient training techniques, offering a balance between performance and computational practicality.

In-context based Multi-task Learning. Learning to perform multiple tasks simultaneously is a longstanding challenge in computer vision. Moving beyond earlier multi-task learning frameworks that depend on fixed task protocols (Doersch & Zisserman, 2017; Sener & Koltun, 2018), there are innovative attempts to integrate various vision tasks under a unified model. For example, Pix2Seq (Chen et al., 2021) pioneered this direction by treating various task outputs as elements in a discrete space. Approaches like Unified-IO (Lu et al., 2022), OFA (Wang et al., 2022), UViM (Kolesnikov et al., 2022), and Painter (Wang et al., 2023b) further streamlined this process by converting diverse inputs and outputs into sequences of tokens.

Building on this, newer methods draw inspiration from in-context learning in large language models, eliminating the task-specific paradigm altogether. These models deduce tasks directly from the input prompt, with (Pathak et al., 2016) predicting missing sections of images and (Bar et al., 2022) merging task examples with a query image to generate outputs via inpainting. LVM (Bai et al., 2023) extends this concept, allowing for an expanded input sample set.

Learning from Long-tail distribution. The long-tailed distribution of sample classes is a prevalent and taxing issue in computer vision (Zhang et al., 2023). Models trained under skewed distributions experience a marked performance drop due to the imbalance in class representation. A straightforward approach to mitigate this issue is to re-balance the train-

ing dataset such that each class is equally represented (Kang et al., 2019). Unlike previous scenarios where imbalances manifest across individual classes, our focus shifts to the disparity in the volume of data across various tasks. As a foundational strategy, we employ simple re-sampling techniques. We further explore traditional data augmentation methods and demonstrate that they not only contribute to a more equitable distribution of training samples among tasks, but also yield compelling results.

Knowledge distillation. Knowledge Distillation (KD) has emerged as a prominent model compression technique over the past decade. The core idea of KD involves leveraging the soft predictions generated by a pre-trained teacher model to guide the training of a more compact student model (Hinton et al., 2015). Additional loss functions have been proposed in subsequent studies to further enhance KD (Zhao et al., 2022; Guo et al., 2024; Hao et al., 2024). Beyond focusing solely on the final output, KD can integrate intermediate features through techniques such as pixel-level matching (Romero et al., 2014; Zhang et al., 2020), attention alignment (Komodakis & Zagoruyko, 2017), relation matching (Peng et al., 2019; Park et al., 2019; Guo et al., 2021), or contrastive learning (Tian et al., 2020). Several works also employ multiple teacher models for distillation (Hao et al., 2022; Zhu et al., 2018). While the success of KD has been validated across various applications, its applicability to LVMs remains relatively unexplored. Our work fills this gap and confirms the efficacy of KD within the domain of LVMs.

8. Conclusion and Discussion

This paper researches on developing efficient autoregression-based general-purpose vision model, oriented towards the over-reliance on colossal models and extensive balanced data. Our observation of impaired results when the model faces unevenly-distributed data across tasks underpins our data augmentation strategy to automatically enrich the limited datasets and boost the overall performance of LVMs. Furthermore, our findings demonstrate the potential of KD to bridge performance and efficiency for autoregressive LVMs. This work serves to advance understanding into autoregressive LVMs and provide a basis for designing more efficient generalist vision models.

Limitation and future direction. While it is now possible to generate outputs for various vision tasks using corresponding prompts at test time, the conversion of visual results into quantifiable outputs warrants further exploration. For instance, post-processing techniques can generate foreground segmentation outputs, and in theory, color mapping can produce COCO-pose compatible outputs for keypoint detection. However, when faced with more complex tasks, it is difficult to generate quantifiable outputs from output

images. Addressing this challenge represents a valuable research direction. Finding effective ways to derive quantifiable outputs for different tasks from VQGAN-decoded images is an area that requires further study.

In addition, our model can generate reasonable output only for those specific tasks it was trained on. For example, when trained with dataset including segmentation, video, pose estimation, and de-raining, our model can only generate resonable output on such tasks. Corresponding examples are shown in the Figure 16 and Figure 17 in appendix. We observed that the model can handle multiple tasks but only when similar tasks are present in the training data.

Acknowledgements

Impact Statement

Given the usage of data augmentation, researchers can leverage existing benchmark datasets from various tasks to conduct LVM-related experiments, eliminating the need for additional efforts to collect high-quality data. This approach can greatly benefit the community in advancing research on general models. Moreover, since vision data can be transformed into "word sentences" and processed using AR-based architectures, it is possible that similar methods can be applied to data from other domains, potentially facilitating the exploration of unified AR models across different fields. However, while the usage of data augmentation helps improve model performance, it also entails potential risks. There are concerns regarding overfitting and bias, particularly when augmentation is overly aggressive and fails to accurately represent real-world scenarios. It also increases the consumption of computational resources. Furthermore, certain augmentation techniques may inadvertently heighten the model's vulnerabilities to specific adversarial attacks.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A., Darrell,

T., Malik, J., and Efros, A. A. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023.

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., and Efros, A. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020.

Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W. T., Rubinstein, M., et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *International conference on machine learning*, 2020.

Chen, T., Saxena, S., Li, L., Fleet, D. J., and Hinton, G. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 2023.

Dai, A. M. and Le, Q. V. Semi-supervised sequence learning. *Advances in neural information processing systems*, 2015.

Doersch, C. and Zisserman, A. Multi-task self-supervised visual learning. In *Proceedings of the IEEE international conference on computer vision*, 2017.

El-Nouby, A., Klein, M., Zhai, S., Bautista, M. A., Toshev, A., Shankar, V., Susskind, J. M., and Joulin, A. Scalable pre-training of large autoregressive image models. *arXiv preprint arXiv:2401.08541*, 2024.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., and Paisley, J. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017a.

Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., and Paisley, J. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3855–3863, 2017b.

Guo, J., Han, K., Wang, Y., Wu, H., Chen, X., Xu, C., and Xu, C. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., and Xu, C. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Guo, J., Chen, H., Wang, C., Han, K., Xu, C., and Wang, Y. Vision superalignment: Weak-to-strong generalization for vision foundation models. *arXiv preprint arXiv:2402.03749*, 2024.

Hao, Z., Luo, Y., Wang, Z., Hu, H., and An, J. Cdfkd-mfs: Collaborative data-free knowledge distillation via multi-level feature sharing. *IEEE Transactions on Multimedia*, 2022.

Hao, Z., Guo, J., Han, K., Hu, H., Xu, C., and Wang, Y. Vanillakd: Revisit the power of vanilla knowledge distillation from small scale to large scale. *arXiv preprint arXiv:2305.15781*, 2023.

Hao, Z., Guo, J., Han, K., Tang, Y., Hu, H., Wang, Y., and Xu, C. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems*, 2024.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Hinton, G. E., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv: 1503.02531*, 2015.

Jiang, K., Wang, Z., Yi, P., Chen, C., Huang, B., Luo, Y., Ma, J., and Jiang, J. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8346–8355, 2020.

Johnson, S. and Everingham, M. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of Computer Vision and Pattern Recognition (CVPR) 2011*, 2011.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Kolesnikov, A., Zhai, X., and Beyer, L. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

Kolesnikov, A., Susano Pinto, A., Beyer, L., Zhai, X., Harmsen, J., and Houlsby, N. Uvim: A unified modeling approach for vision with learned guiding codes. *Advances in Neural Information Processing Systems*, 2022.

Komodakis, N. and Zagoruyko, S. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014.

Lu, J., Clark, C., Zellers, R., Mottaghi, R., and Kembhavi, A. Unified-io: A unified model for vision, language, and multimodal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

Park, W., Kim, D., Lu, Y., and Cho, M. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. Image transformer. In *International conference on machine learning*, 2018.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., and Zhang, Z. Correlation congruence for knowledge distillation. In *International Conference on Computer Vision*, 2019.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 2022.

Sener, O. and Koltun, V. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 2018.

Shaban, A., Bansal, S., Liu, Z., Essa, I., and Boots, B. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

Team, I. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.

Tian, Y., Krishnan, D., and Isola, P. Contrastive representation distillation. In *IEEE/CVF International Conference on Learning Representations*, 2020.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Uria, B., Murray, I., and Larochelle, H. Rnade: The real-valued neural autoregressive density-estimator. *Advances in Neural Information Processing Systems*, 2013.

Van Den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. In *International conference on machine learning*, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 2017.

Wang, J., Zhang, S., Liu, Y., Wu, T., Yang, Y., Liu, X., Chen, K., Luo, P., and Lin, D. Riformer: Keep your vision backbone effective but removing token mixer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, 2022.

Wang, X., Wang, W., Cao, Y., Shen, C., and Huang, T. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Yang, W., Tan, R. T., Feng, J., Liu, J., Guo, Z., and Yan, S. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Yu, Y., Zhan, F., Wu, R., Pan, J., Cui, K., Lu, S., Ma, F., Xie, X., and Miao, C. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

Zhang, L., Shi, Y., Shi, Z., Ma, K., and Bao, C. Task-oriented feature distillation. In *Advances in Neural Information Processing Systems*, 2020.

Zhang, S.-H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., and Hu, S.-M. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

Zhang, Y., Kang, B., Hooi, B., Yan, S., and Feng, J. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Zhao, B., Cui, Q., Song, R., Qiu, Y., and Liang, J. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022.

Zhu, X., Gong, S., et al. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 2018.

A. Training Configurations

A.1. Model architecture

In our main paper, we employ three LVMs with different parameters. These models are structured based on LLaMA (Touvron et al., 2023a). The detailed architecture configurations are outlined in Table 7.

Table 7: Detailed configurations of used models.

Model	Hidden dim.	MLP dim.	#heads	#layers
LLaMA-1B	2048	5504	16	22
LLaMA-300M	1024	2688	8	22
LLaMA-80M	768	3072	12	12

A.2. Training details

Our training strategy adheres to the implementation of LVM (Bai et al., 2023), with slight adjustments made for efficient training with 8-16 A100 GPUs. Our models are trained based on the InternLM framework (Team, 2023). The optimization details are summarized in Table 8. Table 9 presents the time consumption and memory requirements for training each model.

Table 8: Detailed configurations for training efficient LVMs. We attain a consistent equivalent batch size across different models by adjusting the number of employed GPUs, mini-batch size, and gradient accumulation steps.

Config	Value
optimizer	AdamW
learning rate	1.5e-4
weight decay	0.1
optimizer momentum	β_1, β_2=0.9, 0.95
equivalent batch size (tokens)	262144
learning rate schedule	cosine
warmup steps	#total steps * 0.0056
final learning rate	1.5e-5
context length	2048
data augmentation	RandomResizedCrop

Table 9: Training time and memory requirements of each LVM. The equivalent batch size is calculated as the product of the number of utilized GPUs, mini-batch size, and gradient accumulation steps, resulting in a uniform value of 262144 across various models.

Model	Teacher	#GPUs	Mini batch (tokens)	#gradient accum.	Time (hours)	Memory (GB)
LLaMA-1B	-	16	32768	4	324	70
LLaMA-300M	-	16	65536	2	126	70
LLaMA-300M	LLaMA-1B	16	65536	2	235	80
LLaMA-80M	-	8	131072	2	65	69
LLaMA-80M	LLaMA-1B	8	65536	4	82	49

A.3. Dataset

Our full training involves diverse datasets, including Rain13K, SA-1B, COCO-pose, HDvila-100m, and LAION datasets. To ensure data balance, data augmentation is applied to extend the Rain13K and COCO-pose datasets.

Rain13K (16.96%; 14.02 billion tokens). Rain13K serves as the most commonly utilized training dataset for rain removal, consisting of five distinct rain removal datasets. The original dataset comprises 13,712 clean-rain image pairs. We filtered the data and generated a subset for training.

SA-1B (35.05%; 28.97 billion tokens). SA-1B is a large-scale multimodal dataset designed for training general-purpose object

segmentation models. It comprises 1.1 billion high-resolution, diverse, and privacy-protected images, along with corresponding high-quality segmentation masks. We selected a portion of the data for training.

COCO-Pose (12.92%; 10.68 billion tokens). COCO-Pose is a dataset for human pose detection. The original version is a subset of COCO dataset and has 250K images, with annotations for 17 human keypoints, such as eyes, hands, legs, foots, etc. COCO keypoints dataset can be used to train and evaluate various human pose detection models.

HDvila-100m (1.25%; 1.03 billion tokens). HDvila-100m is a large-scale video-language multimodal dataset containing 100 million high-resolution, diverse, and video clips and 100 million automatically generated text descriptions. The dataset covers a wide range of topics and includes both video segments and corresponding text captions. The video segments are 10 seconds long, the text captions are produced by an automatic speech recognition system. We exclusively utilized a subset of the video data to endow our model with the capability for continuous inference.

LAION-400M (33.83%; 27.96 billion tokens). LAION-400M is a large-scale multimodal dataset containing 400 million English image-text pairs. It covers a wide range of topics and includes both text descriptions and corresponding images. We adopt a subset of this dataset in our experiments.

B. Continual Learning.

B.1. Additional results

In our ablation study on data shuffling, we observe catastrophic forgetting in LVMs within an ordered data setting. To further investigate, we assess the impact of rescaling the learning rate at the beginning of each task, a configuration more aligned with the standard continual learning setting. The summarized training configurations are presented in Table 10.

Table 11 presents quantitative results for both scenarios—without and with learning rate rescaling—while Figure 7 showcases the corresponding visualizations. In both settings, the results indicate the presence of catastrophic forgetting in LVMs during continual learning scenarios. Moreover, the adoption of learning rate rescaling results in improved performance on the last-trained task but exacerbates forgetting of other tasks.

Table 10: Learning rate configuration for the continual learning setting. The setup labeled as "w/o learning rate rescaling" corresponds to the experiment detailed in the main paper.

	Learning rate scale at task #		
Learning rate configuration	Task 1 (SA-1B)	Task 2 (COCO-Pose)	Task 3 (Rain13K)
w/o learning rate rescaling	1.5e-4∼8e-5	8e-5∼5e-5	5e-5∼1.5e-5
w/ learning rate rescaling	1.5e-4∼1.5e-5	1.5e-4∼1.5e-5	1.5e-4∼1.5e-5

Table 11: Performance of LVMs in continual learning scenarios. We train LLaMA-300M models without shuffling the training data. Two distinct learning rate scheduling schemes were employed. Under each scheme, the LVM exhibits signs of catastrophic forgetting.

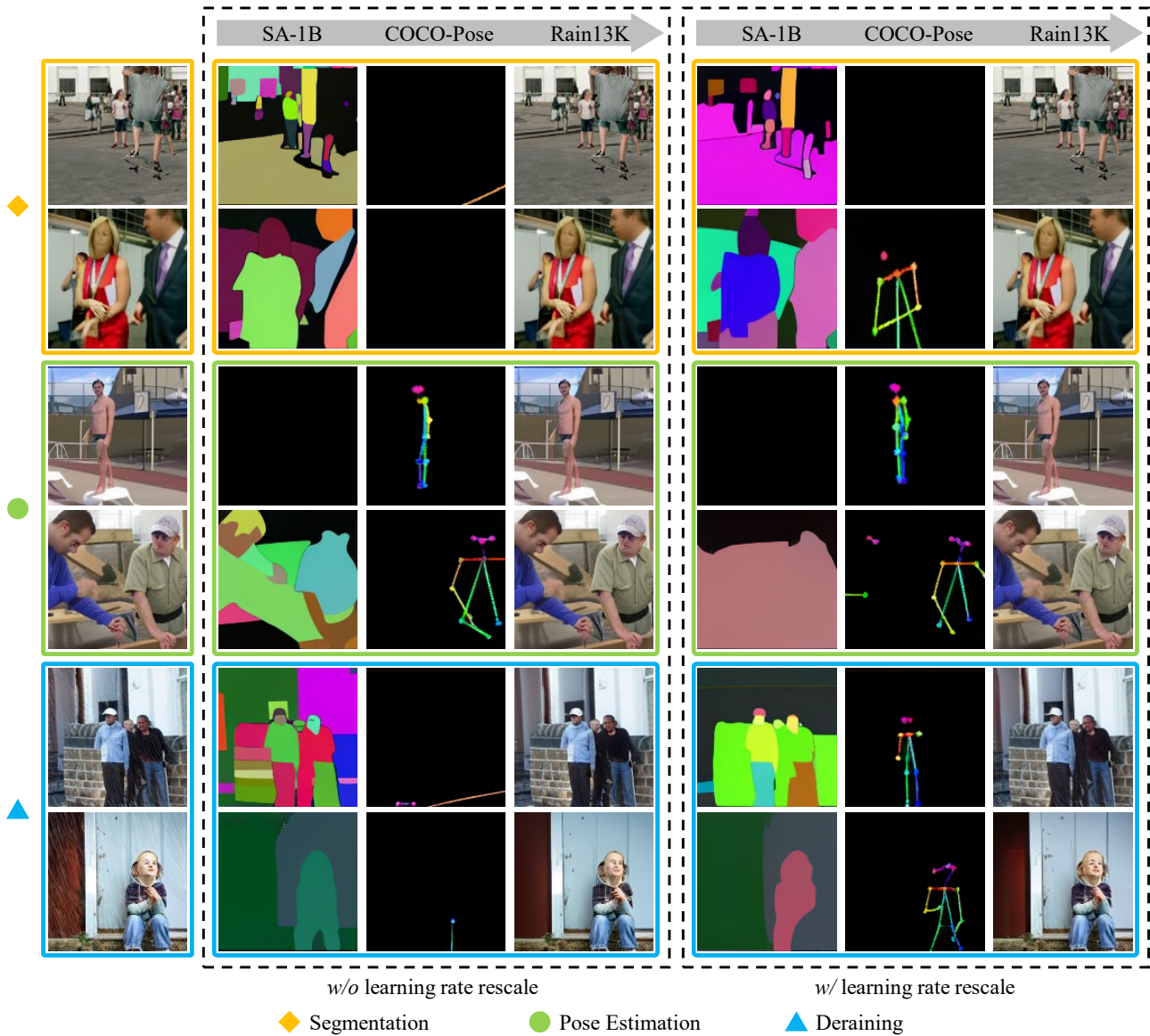|	**Loss ↓**			**Accuracy ↑**			**perplexity ↓**		
Data	Segmentation	Pose Est.	Deraining	Segmentation	Pose Est.	Deraining	Segmentation	Pose Est.	Deraining
w/o learning rate rescaling									
SA-1B	4.62	5.49	7.07	19.17	16.54	1.42	102.06	243.09	1165.21
COCO-Pose	6.02	4.89	7.88	10.29	20.69	0.57	416.21	134.12	2633.08
Rain13K	7.28	7.16	5.70	4.45	5.99	11.79	1449.42	1285.80	287.79
w/ learning rate rescaling									
SA-1B	4.62	5.48	7.11	19.29	16.65	1.34	101.52	241.59	1213.27
COCO-Pose	6.22	4.88	7.98	9.23	20.84	0.54	510.03	132.78	2880.20
Rain13K	7.98	7.78	5.67	2.88	3.92	11.91	2925.92	2403.03	281.81

Figure 7: Visualization of inference results of model at various training stages. Catastrophic forgetting is evident in scenarios with and without learning rate rescaling.

B.2. Offline training

In Figure 8, we compare the inference results of the offline trained LLaMA-1B model on the entire dataset using different prompts but the same inputs. In contrast to the continual learning scenario, the model trained with shuffled data demonstrates successful recognition of the given prompts. These results underscore the ability of the model to adeptly handle multi-task scenarios with distinct prompts.
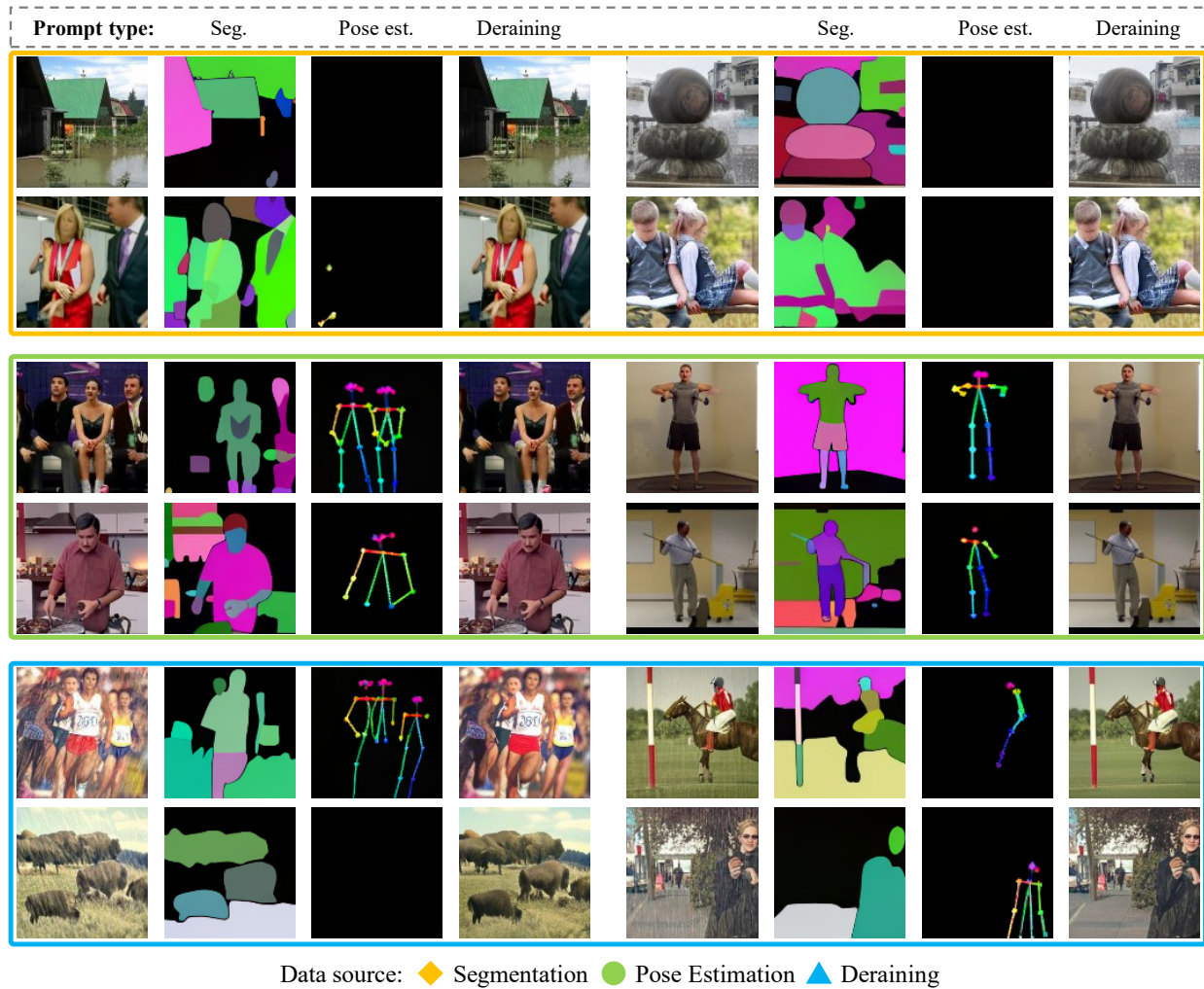

Figure 8: Visualization of inference results of LLaMA-1B with different prompts on the same inputs.

C. Full Results of LLaMA-80M.

We train an LLaMA-80M model employing KD and present its performance in terms of perplexity in the main paper.

Figure 9 provides additional visualization of the generated images. The efficient 80M model exhibits significant potential in handling prompted image autoregressive tasks, and the performance is further enhanced when KD is applied. We posit that by incorporating more high-quality data and extending the training schedule, these efficient models can achieve practical success in the future.

Table 12: Comparison of validation performance between LLaMA-80M models trained with and without KD.

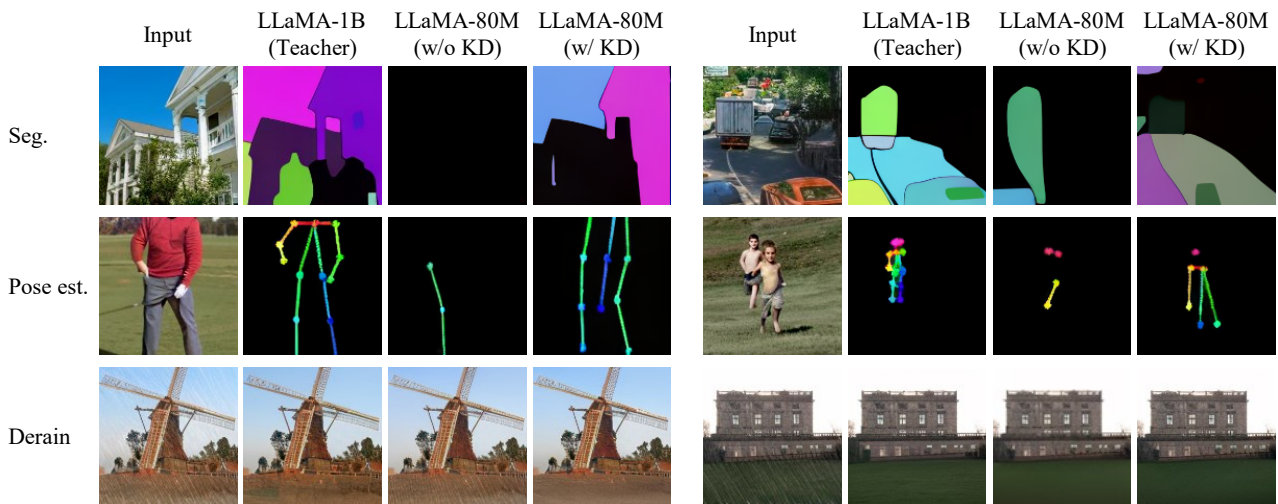| Data	KD	Loss ↓			Accuracy ↑			perplexity ↓		
		Segmentation	Pose Est.	Deraining	Segmentation	Pose Est.	Deraining	Segmentation	Pose Est.	Deraining
LLaMA-1B	-	4.48	4.71	5.47	19.89	21.45	12.46	88.33	111.95	230.94
LLaMA-80M	✗	5.04	5.10	5.60	16.08	19.03	12.05	156.19	164.57	264.56
LLaMA-80M	✓	4.99	5.07	5.57	16.30	19.04	12.13	147.78	159.78	256.65

Figure 9: Visualization of inference results from the efficient LLaMA-80M model.

D. Single-image Inpainting

Figure 10 illustrates the visualization of single-image inpainting. For each input sequence, comprised of 256 tokens representing a single image, the last 128 tokens are removed. Pretrained LVMs are then employed to predict the removed tokens. Subsequently, the generated tokens are concatenated with the initial 128 tokens, and the VQGAN decoder is utilized to obtain the inpainted image.
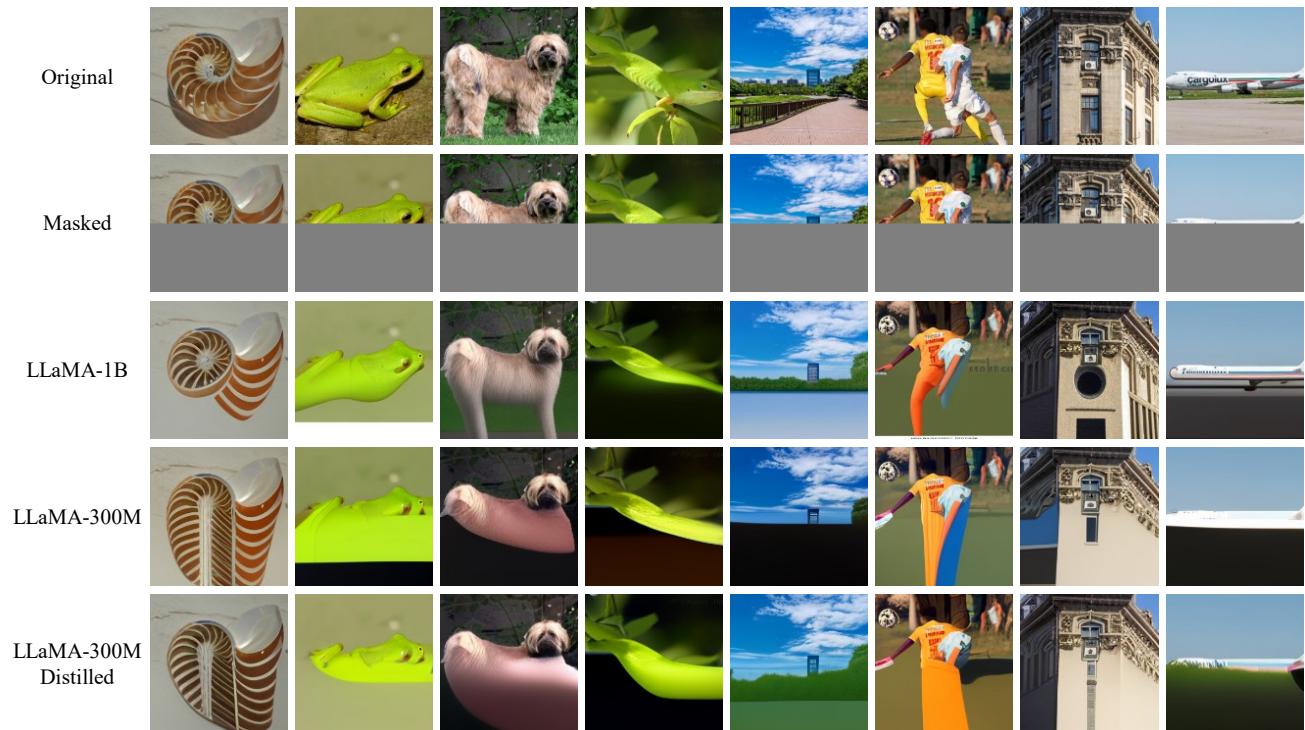

Figure 10: Visualization of image inpainting results. Input of the top half of each image to LVMs to generate the inpainted bottom half.

E. Additional results of KD

We provide more results using the LLaMA-300M models trained with and without KD, which corresponds to the same model presented in Table 3 of the main paper. The evaluation includes pose estimation datasets, specifically Leeds Sports Pose (LSP) (Johnson & Everingham, 2011) in Figure 11, and Occluded Human (OCHuman) (Zhang et al., 2019) in Figure 12. Additionally, we present testing results on the deraining task using rain100H (Yang et al., 2017) in Figures 13 and Figures 14, and rain1400 (Fu et al., 2017a) in Figure 15.

Figure 11: Visualization on the LSP (Leeds Sports Pose) dataset. LLaMA-300M model trained with KD (the third column) obtains better results compared to the model trained without KD.

Query   w/o KD   w/ KD   Query   w/o KD   w/ KD   Query   w/o KD   w/ KD

Figure 12: Visualization on the OCHuman (Occluded Human) dataset. LLaMA-300M model trained with KD (the third column) obtains better results compared to the model trained without KD.

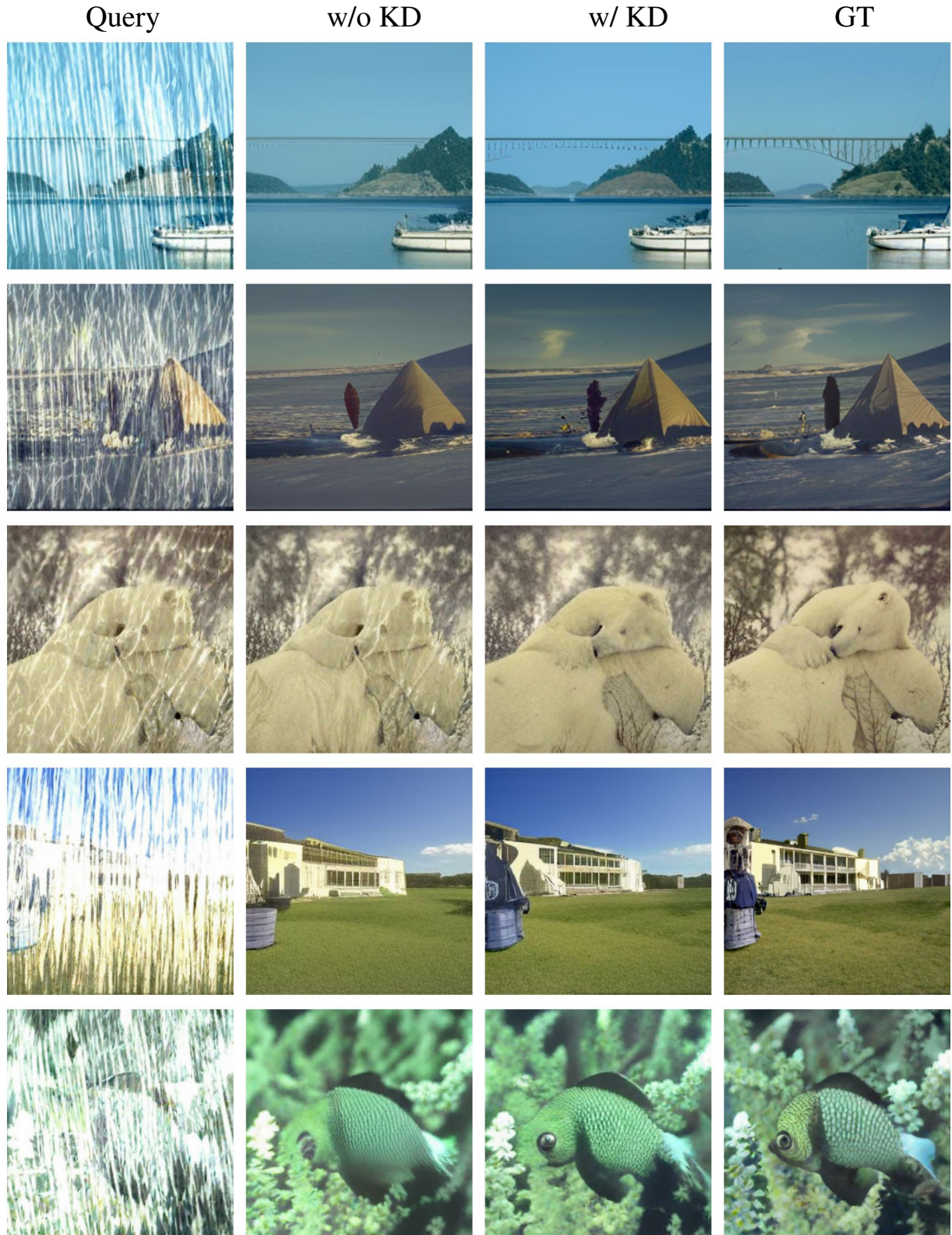| Query	w/o KD	w/ KD	GT

Figure 13: Visualization on the rain100H dataset. LLaMA-300M model trained with KD (the third column) obtains exhibits greater similarity to the ground truth compared to the model trained without KD.

Query	w/o KD	w/ KD	GT

Figure 14: Visualization on the rain100H dataset. LLaMA-300M model trained with KD (the third column) obtains exhibits greater similarity to the ground truth compared to the model trained without KD.

Query	w/o KD	w/ KD	GT

Figure 15: Visualization on the rain1400 dataset. LLaMA-300M model trained with KD (the third column) obtains exhibits greater similarity to the ground truth compared to the model trained without KD.
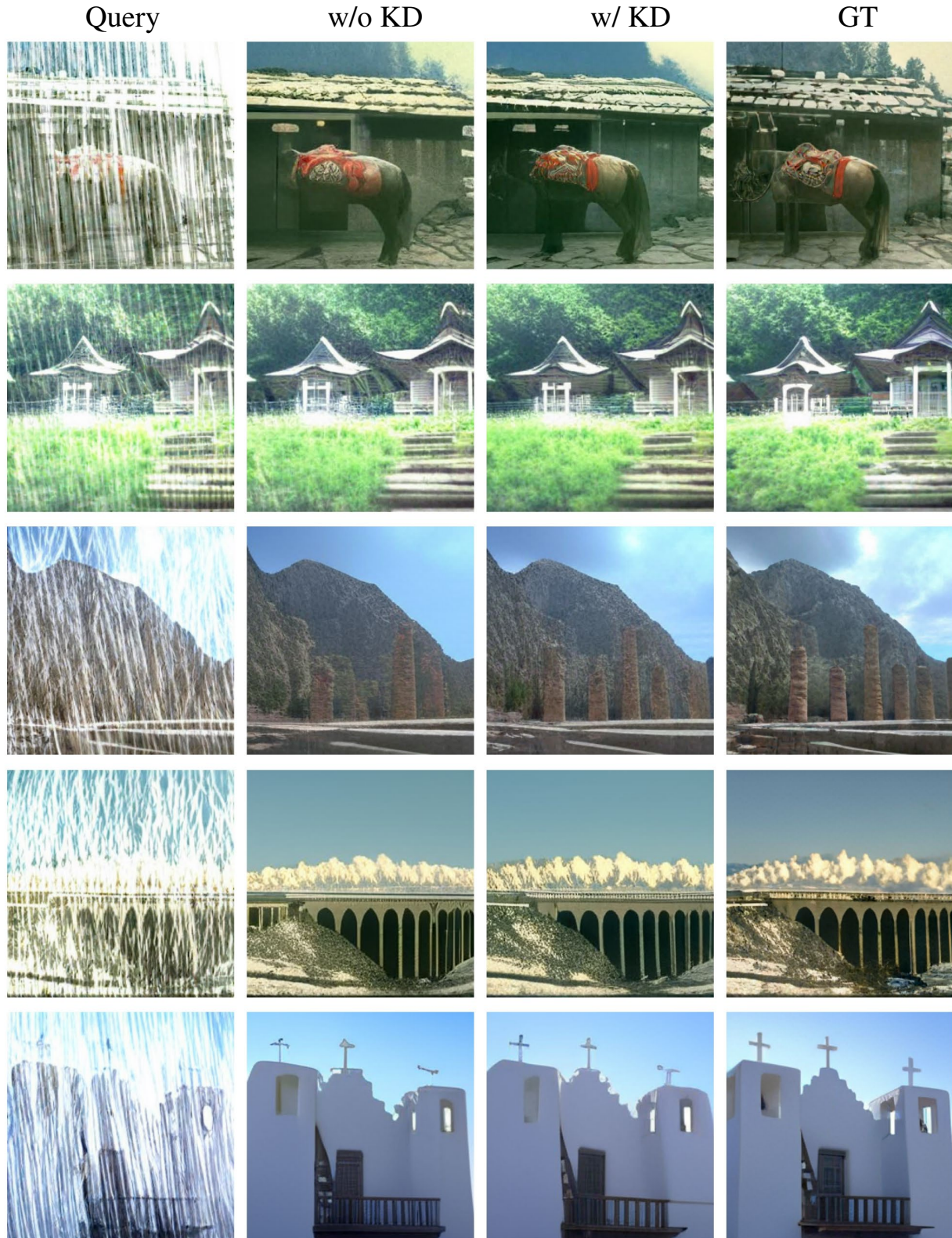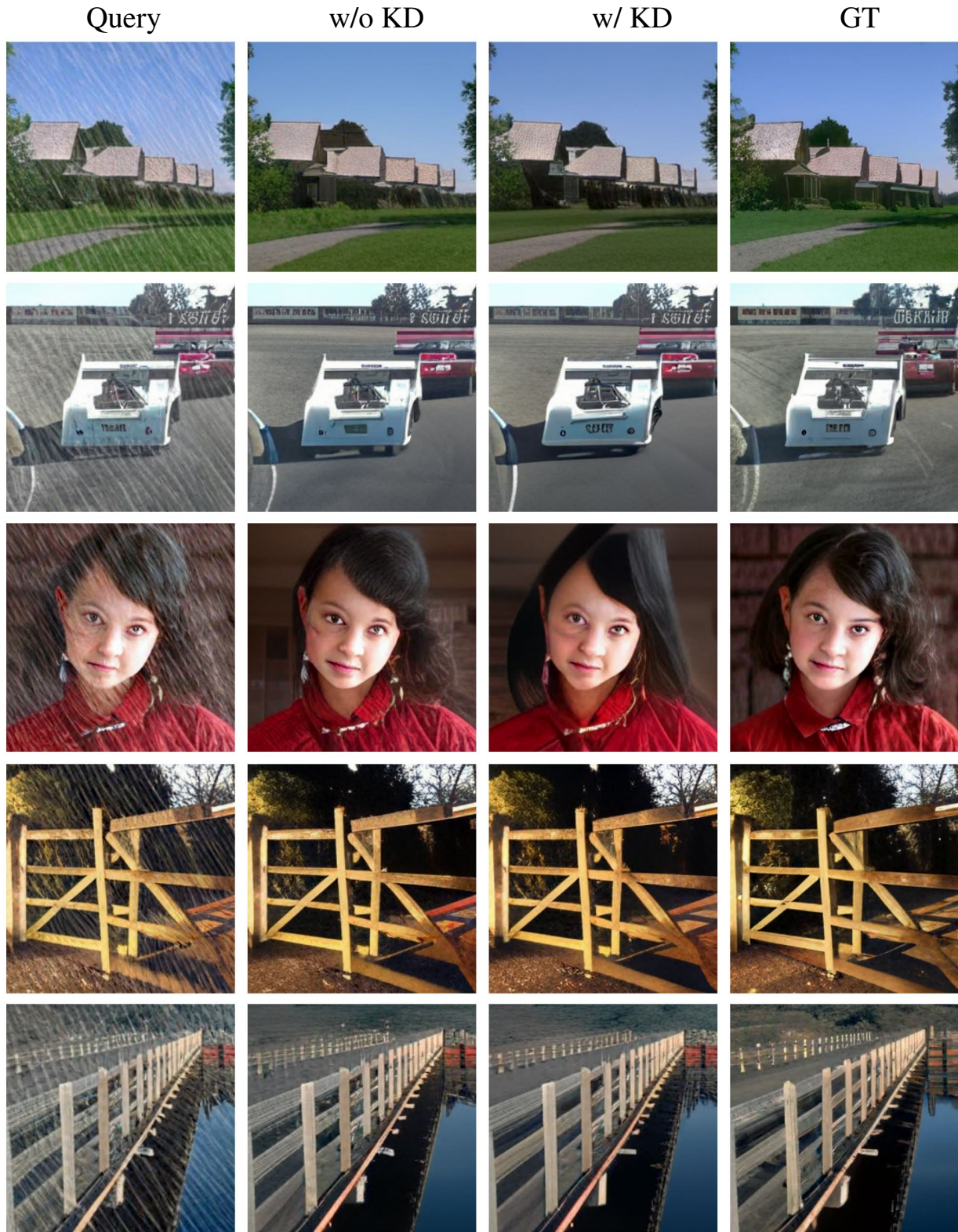
Prompt   Output

Frame predictions: considering dynamic objects and camera motion.

Frame predictions: zoom in.

Frame predictions: zoom out.

Figure 16: The HD-VILA-100M video data is integrated into the training process of our LLaMA-1B, enabling it to handle frame predictions.

Prompt   Output

Depth estimation: image to depth.

Inpainting: partially masked image to orignial image.

Colorization: gray-scale image to color-scale image.
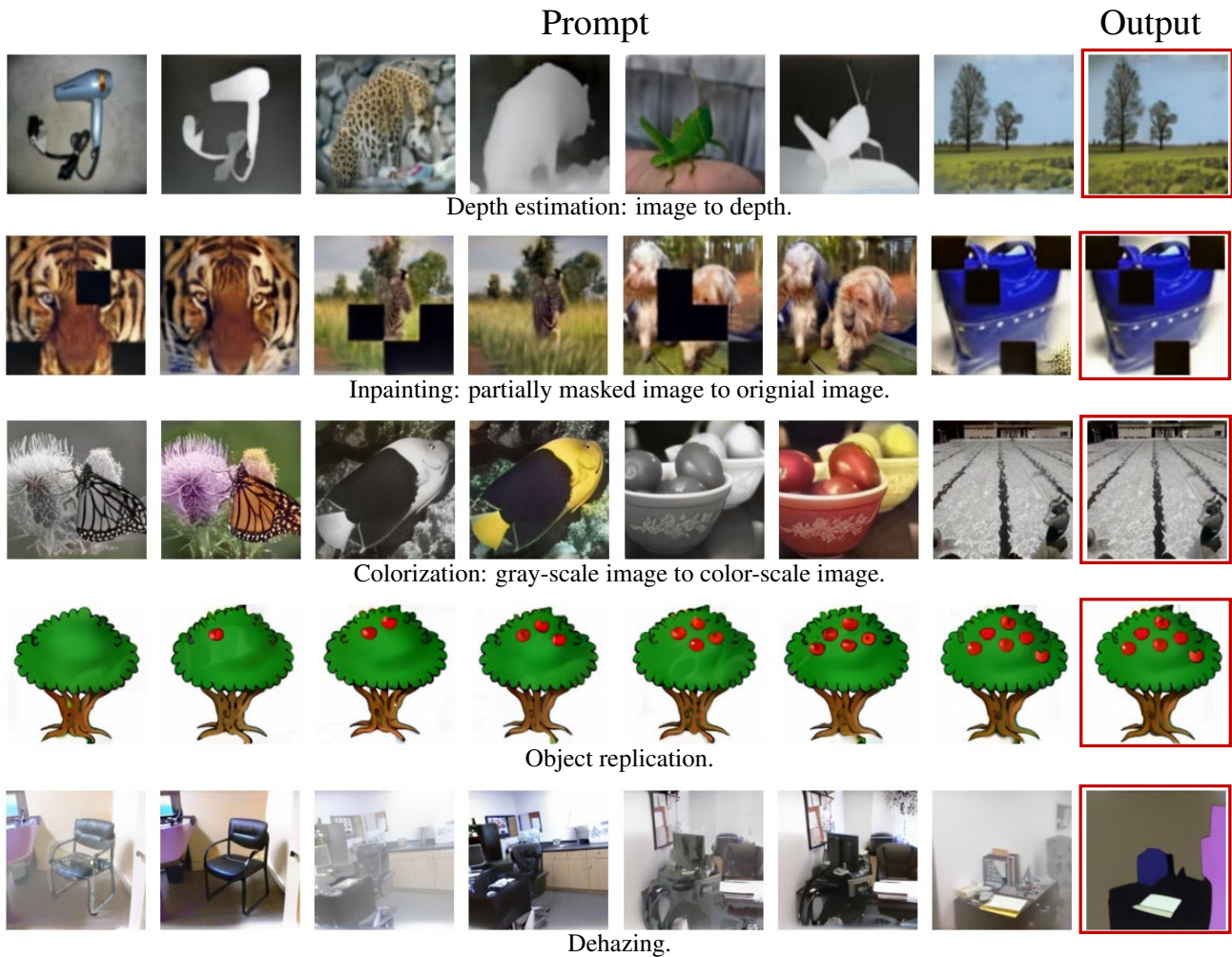
Object replication.

Dehazing.

Figure 17: For tasks that are not included in its training, the LLaMA-1B struggles to generate satisfactory output and resorts to repeating the last available image. For the novel dehazing task, it was mistakenly interpreted as a segmentation task.