Towards Understanding and Quantifying Uncertainty for Text-to-Image Generation

Supplementary Material

A. Prompt similarity score

ROUGE [36] and BERTScore [72] are conceptually similar scores that measure how well a candidate text sequence (caption in our case) matches a reference text sequence (prompt). ROUGE operates at the text/token level whilst BERTScore is calculated on the deep embeddings of a pre-trained text encoder. Concretely they are calculated as follows:

BERTScore

$$\text{Precision} = \frac{1}{|C|}\sum_{c\in C}\max_{r\in R}\cos(c,r) \tag{4}$$

$$\text{Recall} = \frac{1}{|R|}\sum_{r\in R}\max_{c\in C}\cos(r,c) \tag{5}$$

Where C and R are the sets of token embeddings for the candidate and reference, respectively. $\cos(x, y)$ is the cosine similarity between two embeddings.

ROUGE-n

$$\text{Precision} = \frac{\sum_{g\in G(C)}\min(\text{count}(g,C),\text{count}(g,R))}{\sum_{g\in G(C)}\text{count}(g,C)} \tag{6}$$

$$\text{Recall} = \frac{\sum_{g\in G(C)}\min(\text{count}(g,C),\text{count}(g,R))}{\sum_{g\in G(R)}\text{count}(g,R)} \tag{7}$$

Where $G(C)$ and $G(R)$ are the sets of n-grams for the candidate and reference, respectively. $\text{count}(g, S)$ is the count of n-gram g in sequence S.

ROUGE-L

$$\text{Precision} = \frac{\text{LCS}(C,R)}{|C|} \tag{8}$$

$$\text{Recall} = \frac{\text{LCS}(C,R)}{|R|} \tag{9}$$

Where $\text{LCS}(C, R)$ is the length of the longest common subsequence between C and R. $|C|$ and $|R|$ are the lengths of the candidate and reference sequences.

Comparing to the ideas of semantic-concept-based recall and precision presented in Sec. 4.2, we can see how the above formulae approximate concepts with *e.g.* n-grams, BERT embeddings. Eqs. (6) and (7) directly substitute semantic concepts with n-grams for example, allowing for the approximate quantification of precision and recall based semantic uncertainty.

B. Extra Results

Is PUNC BERT (precision) more effective than PUNC BERT (recall) for quantifying aleatoric uncertainty? And conversely, is PUNC BERT (recall) better suited for quantifying epistemic uncertainty? Table 5 presents the recall-to-precision ratio, calculated as $\frac{\text{PUNC BERT (recall)}}{\text{PUNC BERT (precision)}}$, for epistemic (OOD) and aleatoric uncertainty scenarios. For epistemic uncertainty, the ratio exceeds one for both AUPR and AUROC, favoring recall, which prioritizes identifying true positives even at the cost of higher false positives. In contrast, aleatoric uncertainty exhibits ratios below one, favoring precision and minimizing false positives despite lower recall. These trends are supported by FPR95 values, with lower rates for epistemic uncertainty, indicating better OOD separation, and higher rates for aleatoric uncertainty, reflecting challenges in handling ambiguous inputs. This analysis underscores the importance of tailoring uncertainty quantification strategies to the specific nature of the uncertainty—epistemic or aleatoric—to optimize model performance.

	Epistemic			Aleatoric			
	Microscopic	Remote Sensing	Texture	Vague	Adversarial	Corrupt lvl1	Corrupt lvl2
auroc ↑	2.25	1.65	1.12	0.06	0.30	0.45	0.44
aupr ↑	1.88	1.56	1.05	0.83	0.95	0.57	0.57
fpr95 ↓	0.44	0.78	0.94	1.00	1.16	1.06	1.16

Table 5. $\frac{\text{PUNC BERT (recall)}}{\text{PUNC BERT (precision)}}$ performance ratio using Molmo LVLM.

What is the best measure of similarity at the image level? Identifying the most appropriate similarity metric at the image level is non-trivial. To better understand this, we have included additional results in Tables 7 and 8. These results suggest that the LPIPS metric [71] often yields the best performance. However, in some cases, combining LPIPS with MSE leads to improved results, while in others, MSE alone performs best. This variation largely depends on the specific uncertainty dataset being used.

How robust is PUNC to changes in the underlying LVLM? To evaluate the robustness of PUNC with respect to different Large Vision-Language Models (LVLMs), we provide results in Tables 7 and 8 using LLAVA Next and LLaMA 3.2, and in Table 11 using QWEN. These results indicate that performance varies significantly depending on the chosen LVLM, which highlights the sensitivity of uncertainty quantification to the underlying model. To mitigate this dependency and improve robustness, we

propose an ensemble-based approach inspired by Deep Ensembles [32] and Test-Time Augmentation. Our method aggregates the BERT Recall and BERT Precision scores from three state-of-the-art VQA models—LLAVA, Molmo, and QWEN—by averaging their outputs. As shown in Table 11, the ensemble of BERT Recall scores consistently outperforms all other evaluated approaches. This demonstrates the effectiveness of our ensemble strategy in enhancing both calibration and robustness in VQA tasks.

C. Experimental settings

In this work, we evaluate the performance and capabilities of PUNC under various experimental conditions, testing several text-to-image (T2I) models: Stable Diffusion 1.5 (SDv1.5) [50], SDXL [48], PixArt-Σ [7], and SDXS [68]. All models are sourced from the Hugging Face repository and used with their respective default configurations, as summarized in Tab. 9.

For inference, SDXS utilizes a single-step generation process, whereas the other models perform 20 inference steps. Regarding the guidance scale, SDXS applies no guidance, PixArt-Σ uses a scale of 4.5, and both SDv1.5 and SDXL employ a scale of 7.5.

To evaluate the influence of various Large Vision-Language Models (LVLMs) and their caption generation capabilities on the performance of PUNC, we conduct experiments using three models: LLAMA3.2 Vision [15], LLava-Next [35], and Molmo [11]. All models are implemented with their default configurations as provided in the Hugging Face library, and their specifications are summarized in Tab. 10.

Each model is sourced directly from the Hugging Face repository and evaluated under its standard settings. For the experiments, we use the maximum token length permissible for each model in relation to the given prompts. This ensures a consistent and comprehensive comparison of their captioning capabilities across the tested scenarios. Note that for SDXL, SDv1.5, and SDXS, the maximum token length is limited to 77 tokens, while for PixArt-Σ, a maximum of 300 tokens is permitted.

D. Ablation Results

This section presents additional experiments as part of our ablation study, focusing on the model choices for PUNC, particularly the selection of Large Vision-Language Models (LVLMs) that form the core of PUNC. The results of these experiments are provided in Table 11 for the epistemic datasets and Table 12 for the aleatoric datasets.

E. Dataset of prompts

We need an In Distribution dataset of prompt and Distribution datasets of prompt with uncertainty. For the In Distribution dataset of prompt we have choose to use the dataset proposed by [13], where the authors generate high-quality descriptions for each training image of ImageNet [12] by interacting with GPT-4. These descriptions include detailed prompt, providing a high information about the image content. We have choosen to use prompt of ImageNet images because this images are use as pretrained of most diffusion model since a lot of them use DiT [46] backbone or a variant of this diffusion model. We denote this dataset of prompt as Normal.

Regarding the Uncertainty prompt, since we have two kind of uncertainty we have mainly to kind to build the prompt. First regarding the OOD dataset of prompt we have used images of remote sensing [47], texture [9], Microscopic [44] datasets. The idea is that diffusion model trained on LAION-5b dataset [51] often lack good perfomences on this dataset due to the fact that this kind of images are quite absent on LAION-5b. Then once the we have collected the images, we used LLAVA next for its efficiency and velocity to caption the images. We denote these dataset of prompt respectively as Remote sensing, Texture, and Microscopic.

Regarding the aleatoric uncertainty prompt, we proposes two datasets. One called vague which is composed of 2k prompts of the following shape: "An image of ***." "An picture of ***." where we replaced the *** with the name of the class of ImageNet. We denote this dataset of prompt as Vague.

To conduct our experiments, we require both an *in-distribution dataset of prompts* and d*atasets of prompts with uncertainty*.

In-Distribution Dataset of Prompts For our in-distribution dataset, we use the dataset proposed by [13], in which the authors generate high-quality descriptions for each training image in the ImageNet dataset [12] by interacting with GPT-4. These descriptions provide detailed and contextually rich prompts that thoroughly represent the image content. We selected randomly 20 by class prompts derived from ImageNet images because many diffusion models are pretrained on ImageNet, given that they often utilize a DiT backbone or variants of this architecture. Using prompts that align with the content on which these models are pretrained ensures the consistency of the in-distribution data. We refer to this dataset as *Normal*.

Out-of-Distribution (OOD) Prompts for Epistemic Uncertainty For prompts representing epistemic uncertainty, we need out-of-distribution (OOD) content, as this type of uncertainty arises when models encounter unfamiliar or untrained data. For this purpose, we selected images from domains typically underrepresented in the LAION dataset [51], which many diffusion models are trained on. These domains include remote sensing, texture, and microscopic

	Microscopic			Remote Sensing			Texture		
	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓
BERT recall (Molmo)	97.83%	98.41%	10.32%	91.93%	93.08%	43.30%	84.40%	95.14%	69.99%
BERT recall (lvlm Ensemble including QWEN)	**98.29%**	**98.77%**	**7.58%**	**94.00%**	**94.90%**	**33.71%**	**91.20%**	**97.53%**	**53.67%**
2XDM mse	35.43%	43.22%	100.00%	56.52%	54.87%	99.26%	54.19%	24.78%	92.23%
2XDM lpips_alex	24.32%	33.21%	99.99%	54.43%	57.99%	98.78%	37.21%	23.36%	99.37%
n-XDM mse (pairwise similarity from $n = 7$ T2I inferences)	32.95%	78.83%	100.00%	64.46%	89.94%	100.00%	55.85%	69.12%	96.13%
n-XDM lpips_alex (pairwise similarity from $n = 7$ T2I inferences)	11.64%	61.71%	100.00%	58.47%	87.85%	99.89%	28.81%	48.68%	99.98%
2XDM mse with CLIP embedding	62.84%	64.53%	98.97%	66.01%	70.45%	88.70%	52.68%	50.08%	97.73%

Table 6. OOD Experiments (please compare with Table 2 of the main paper)

		SDXS			PixArt			SDv1.5			SDXL			Average		
		auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓
microscopic	DDPM-OOD mse				39.61%	56.44%	100.00%	71.87%	71.84%	93.59%	87.39%	83.91%	51.10%	66.29%	70.73%	81.56%
	DDPM-OOD lpips_alex				24.24%	32.53%	100.00%	22.62%	33.04%	99.69%	18.09%	30.85%	99.96%	21.65%	32.14%	99.89%
	DDPM-OOD mse_and_lpips				39.61%	56.38%	100.00%	71.64%	71.52%	93.62%	87.25%	83.67%	51.43%	66.17%	70.53%	81.68%
	LMD mse				41.27%	58.00%	100.00%	72.20%	69.86%	85.56%	55.06%	48.72%	89.80%	56.17%	58.86%	91.79%
	LMD lpips_alex				45.82%	45.24%	99.10%	16.81%	31.61%	99.88%	5.75%	28.12%	100.00%	22.79%	34.99%	99.66%
	2XDM mse	31.02%	35.89%	99.77%	35.43%	43.22%	100.00%	54.21%	52.27%	98.88%	65.66%	52.85%	76.71%	46.58%	46.06%	93.84%
	2XDM lpips_alex	24.49%	32.60%	99.91%	24.32%	33.21%	99.99%	16.19%	32.42%	99.94%	30.58%	36.99%	98.25%	23.90%	33.81%	99.52%
	PUNC ROUGE (1_recall)	84.00%	85.10%	57.07%	95.47%	97.06%	30.84%	87.24%	90.66%	48.90%	83.56%	84.31%	52.78%	87.57%	89.28%	47.40%
	PUNC ROUGE (L_recall)	87.64%	89.56%	55.28%	96.96%	98.02%	16.85%	**90.20%**	**93.40%**	**43.56%**	**88.19%**	**89.84%**	**47.44%**	**90.75%**	**92.71%**	**40.78%**
	PUNC BERT (recall)	**88.45%**	**89.71%**	**48.68%**	**97.83%**	**98.41%**	**10.32%**	87.75%	91.70%	51.03%	83.10%	85.46%	58.14%	89.28%	91.32%	42.04%
remote sensing	DDPM-OOD mse				79.96%	85.31%	99.28%	74.16%	71.64%	71.18%	65.11%	57.88%	72.29%	73.07%	71.61%	80.92%
	DDPM-OOD lpips_alex				76.56%	82.25%	97.47%	**86.11%**	**86.43%**	65.79%	85.10%	86.12%	58.96%	82.59%	84.93%	74.07%
	DDPM-OOD mse_and_lpips				80.00%	85.37%	99.28%	74.36%	71.89%	71.00%	65.47%	58.25%	71.97%	73.28%	71.84%	80.75%
	LMD mse				83.95%	89.53%	99.05%	75.76%	71.88%	66.83%	58.05%	53.17%	85.33%	72.59%	71.53%	83.74%
	LMD lpips_alex				90.65%	93.11%	74.49%	78.93%	79.31%	81.05%	86.60%	85.39%	49.19%	85.39%	85.94%	68.24%
	2XDM mse	60.13%	56.57%	86.40%	56.52%	54.87%	99.26%	28.62%	39.11%	98.44%	9.28%	31.95%	98.99%	38.64%	45.62%	95.77%
	2XDM lpips_alex	53.88%	59.21%	98.23%	54.43%	57.99%	98.78%	30.66%	38.80%	98.76%	23.54%	35.35%	96.23%	40.63%	47.83%	98.00%
	PUNC ROUGE (1_recall)	85.59%	86.59%	63.87%	94.14%	95.97%	49.29%	73.24%	72.93%	79.56%	73.07%	74.54%	83.48%	81.51%	82.51%	69.05%
	PUNC ROUGE (L_recall)	**89.85%**	**90.63%**	**51.96%**	**97.01%**	**97.94%**	**16.08%**	83.63%	84.65%	**65.35%**	**85.68%**	**87.09%**	**64.00%**	**89.04%**	**90.08%**	**49.34%**
	PUNC BERT (recall)	80.50%	82.56%	75.19%	91.93%	93.08%	43.30%	61.27%	64.74%	90.08%	60.71%	65.32%	90.94%	73.60%	76.42%	74.88%
Texture	DDPM-OOD mse				73.27%	62.14%	89.25%	70.33%	56.23%	99.80%	74.33%	52.19%	**84.62%**	72.64%	56.86%	91.22%
	DDPM-OOD lpips_alex				43.94%	27.22%	97.83%	34.70%	19.45%	99.62%	34.37%	22.32%	99.97%	37.67%	22.99%	99.14%
	DDPM-OOD mse_and_lpips				73.22%	62.02%	89.24%	70.21%	55.97%	99.80%	**74.24%**	51.97%	84.71%	**72.56%**	56.65%	91.25%
	LMD mse				81.09%	71.12%	80.52%	**71.44%**	51.24%	99.81%	56.52%	31.20%	95.00%	69.68%	51.19%	91.78%
	LMD lpips_alex				63.40%	50.47%	97.52%	26.74%	16.06%	99.89%	35.66%	23.22%	99.95%	41.93%	29.92%	99.12%
	2XDM mse	**57.17%**	27.12%	97.11%	54.19%	24.78%	92.23%	46.89%	20.82%	99.83%	51.23%	21.29%	96.95%	52.37%	23.50%	96.53%
	2XDM lpips_alex	32.80%	20.84%	99.98%	37.21%	23.36%	99.37%	29.58%	24.27%	99.93%	28.87%	16.49%	98.88%	32.11%	21.24%	99.54%
	PUNC ROUGE (1_recall)	40.77%	72.88%	95.54%	83.26%	94.80%	73.48%	46.61%	75.64%	**90.87%**	49.16%	78.04%	93.57%	54.95%	80.34%	**88.36%**
	PUNC ROUGE (L_recall)	45.64%	76.90%	95.99%	84.10%	95.13%	74.21%	47.60%	77.70%	93.14%	50.71%	**80.05%**	94.35%	57.01%	82.44%	89.42%
	PUNC BERT (recall)	49.11%	**79.11%**	**94.65%**	**84.40%**	**95.14%**	**69.99%**	47.60%	**78.80%**	93.85%	46.72%	79.20%	96.87%	56.96%	**83.06%**	88.84%

Table 7. **Performance comparison of PUNC applied to different Out of Distribution concepts and different text-to-image models**

images, which are relatively rare in the Laion-5b dataset [51], potentially leading to poorer model performance on these types. After collecting images from the test sets of remote sensing [47], texture [9], and microscopic datasets [44], we used the LLAVA Next model [35] to caption each image. LLAVA Next was chosen for its efficiency and speed in generating relevant descriptions for images from diverse, specialized domains. We label these OOD prompt datasets as *Remote Sensing*, *Texture*, and *Microscopic*.

Aleatoric Uncertainty Prompts For prompts that simulate aleatoric uncertainty, which stems from inherent noise or ambiguity in the prompt, we designed two datasets:

1. *Vague*: This dataset contains 2,000 prompts with deliberately vague descriptions, structured to provide minimal context, such as "An image of ***" or "A picture of ***", where "***" is replaced by the ImageNet class name. These prompts create an ambiguous context, simulating scenarios where the input information is too sparse for the model to fully comprehend. We refer to this dataset as *Vague*.
2. *Adversarial*: This dataset contains 1,000 prompts altered from the *Normal* dataset using UnlearnDiffAtk [73], a gradient-based adversarial attack method optimizing adversarial prompts within the diffusion process.
3. *Corrupted*: To simulate real-world scenarios with input noise, we created prompts with grammatical errors and word omissions. Using LLAMA-3-2, we generated

		SDXS			PixArt			SDv1.5			SDXL			Average		
		auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓
Vague	DDPM-OOD mse				31.44%	6.07%	96.45%	50.07%	9.29%	96.56%	40.63%	5.36%	97.34%	40.71%	6.91%	96.78%
	DDPM-OOD lpips_alex				33.60%	6.30%	97.45%	48.08%	10.09%	97.43%	46.95%	6.27%	98.53%	42.88%	7.55%	97.80%
	DDPM-OOD mse_and_lpips				31.41%	6.06%	96.45%	50.06%	9.30%	96.52%	40.59%	5.36%	97.37%	40.69%	6.91%	96.78%
	LMD mse				31.59%	6.09%	97.83%	48.32%	8.84%	95.78%	41.27%	5.72%	97.79%	40.39%	6.88%	97.13%
	LMD lpips_alex				39.26%	6.91%	96.35%	47.62%	9.54%	97.70%	40.77%	5.70%	97.95%	42.55%	7.38%	97.33%
	2XDM mse	82.69%	55.64%	88.32%	44.25%	8.77%	94.35%	51.42%	9.89%	96.01%	56.58%	9.46%	95.69%	58.74%	20.94%	93.59%
	2XDM lpips_alex	51.28%	10.10%	94.35%	43.82%	8.57%	96.89%	50.97%	10.09%	95.96%	57.83%	10.01%	94.37%	50.98%	9.69%	95.39%
	PUNC ROUGE (1_precision)	**99.99%**	**100.00%**	0.00%	**100.00%**	**100.00%**	**0.00%**	**99.99%**	**100.00%**	**0.00%**	**99.99%**	**100.00%**	0.00%	**99.99%**	**100.00%**	0.00%
	PUNC ROUGE (L_precision)	**99.99%**	**100.00%**	0.00%	**100.00%**	**100.00%**	**0.00%**	**99.99%**	**100.00%**	**0.00%**	**100.00%**	**100.00%**	0.00%	**100.00%**	**100.00%**	0.00%
	PUNC BERT (precision)	60.63%	92.24%	82.85%	67.02%	94.54%	83.10%	37.35%	86.51%	92.70%	70.85%	95.13%	78.00%	58.96%	92.11%	84.16%
Adversarial	DDPM-OOD mse				*39.56%*	*1.78%*	*95.72%*	*50.86%*	*2.53%*	*95.92%*	*38.98%*	*1.10%*	*98.38%*	43.14%	1.81%	96.67%
	DDPM-OOD lpips_alex				*39.59%*	*1.81%*	*97.16%*	*53.65%*	*3.40%*	*96.20%*	*50.39%*	*1.42%*	*97.26%*	47.88%	2.21%	96.87%
	DDPM-OOD mse_and_lpips				*39.54%*	*1.78%*	*95.72%*	*50.89%*	*2.54%*	*95.91%*	*38.97%*	*1.10%*	*98.38%*	43.13%	1.81%	96.67%
	LMD mse				*40.74%*	*1.83%*	*97.41%*	*46.95%*	*2.27%*	*96.35%*	*37.90%*	*1.09%*	*98.80%*	41.87%	1.73%	97.52%
	LMD lpips_alex				*43.71%*	*1.95%*	*96.24%*	*53.29%*	*2.89%*	*94.43%*	*41.64%*	*1.17%*	*97.48%*	46.21%	2.00%	96.05%
	2XDM mse	*58.34%*	*12.65%*	*98.86%*	*49.40%*	*2.45%*	*93.75%*	*53.12%*	*2.75%*	*96.46%*	*57.46%*	*2.09%*	*92.32%*	54.58%	4.99%	95.35%
	2XDM lpips_alex	*36.94%*	*1.94%*	*98.92%*	*47.95%*	*2.54%*	*96.03%*	*53.04%*	*2.83%*	*95.69%*	*55.19%*	*1.96%*	*96.20%*	48.28%	2.32%	96.71%
	PUNC ROUGE (1_precision)	*64.14%*	*97.77%*	*48.80%*	*76.32%*	*98.90%*	*48.00%*	*63.59%*	*97.73%*	*49.50%*	*62.97%*	*97.64%*	*49.20%*	66.76%	98.01%	48.87%
	PUNC ROUGE (L_precision)	*63.20%*	*97.75%*	*49.00%*	*76.26%*	*98.90%*	*48.40%*	*64.09%*	*97.90%*	***49.70%***	*63.76%*	*97.87%*	*49.60%*	66.83%	98.11%	49.17%
	PUNC BERT (precision)	*73.98%*	*98.72%*	*78.31%*	*72.95%*	*98.93%*	*86.40%*	*45.09%*	*96.93%*	*94.18%*	*72.69%*	*98.80%*	*81.40%*	66.18%	98.34%	85.07%
Corrupt lvl1	DDPM-OOD mse				48.89%	44.83%	87.43%	53.15%	54.39%	95.48%	41.01%	29.53%	96.34%	47.68%	42.92%	93.08%
	DDPM-OOD lpips_alex				47.05%	45.74%	89.31%	**58.30%**	**58.59%**	93.56%	53.71%	37.37%	92.42%	53.02%	47.23%	91.76%
	DDPM-OOD mse_and_lpips				48.87%	44.82%	87.43%	53.20%	54.43%	95.48%	41.02%	29.54%	96.32%	47.70%	42.93%	93.08%
	LMD mse				51.34%	45.94%	86.93%	49.03%	51.08%	96.32%	34.80%	27.24%	98.14%	45.06%	41.42%	93.80%
	LMD lpips_alex				51.23%	47.48%	89.99%	57.53%	56.49%	90.92%	40.94%	30.30%	96.32%	49.90%	44.76%	92.41%
	2XDM mse	33.98%	44.40%	99.33%	57.04%	54.03%	87.71%	53.23%	52.17%	95.62%	52.80%	37.94%	94.55%	49.26%	47.14%	94.30%
	2XDM lpips_alex	25.48%	36.16%	99.29%	50.43%	49.86%	90.10%	54.97%	54.29%	95.69%	54.73%	38.27%	93.29%	46.40%	44.65%	94.59%
	PUNC ROUGE (1_precision)	28.75%	38.34%	98.77%	54.58%	55.29%	96.02%	27.69%	37.57%	98.78%	27.45%	37.27%	98.54%	34.62%	42.12%	98.03%
	PUNC ROUGE (L_precision)	27.29%	38.11%	99.15%	54.32%	55.59%	96.16%	27.30%	38.07%	99.45%	27.35%	37.89%	99.37%	34.07%	42.42%	98.53%
	PUNC BERT (precision)	**85.85%**	**84.45%**	**73.94%**	**79.43%**	**78.83%**	87.21%	**54.97%**	**53.73%**	94.51%	**75.76%**	**72.91%**	84.65%	**74.00%**	**72.48%**	**85.08%**
Corrupt lvl2	DDPM-OOD mse				*35.94%*	*39.33%*	*92.36%*	*54.47%*	*56.66%*	*96.21%*	*37.25%*	*28.45%*	*97.12%*	42.55%	41.48%	95.23%
	DDPM-OOD lpips_alex				*43.23%*	*43.94%*	*91.23%*	*59.46%*	*60.14%*	*93.71%*	*48.79%*	*34.44%*	*95.74%*	50.49%	46.17%	93.56%
	LMD mse				*39.88%*	*41.02%*	*92.72%*	*50.48%*	*52.68%*	*96.59%*	*40.28%*	*31.01%*	*97.97%*	43.55%	41.57%	95.76%
	LMD lpips_alex				*46.90%*	*45.65%*	*92.04%*	***59.94%***	***58.90%***	***90.54%***	*40.88%*	*30.42%*	*96.35%*	49.24%	44.99%	92.98%
	2XDM mse	*55.88%*	*59.72%*	*95.30%*	*51.98%*	*52.68%*	*90.13%*	*53.08%*	*53.10%*	*96.74%*	*55.06%*	*40.98%*	*94.80%*	54.00%	51.62%	94.24%
	2XDM lpips_alex	*58.15%*	*58.50%*	*93.98%*	*52.08%*	*53.36%*	*90.19%*	*53.57%*	*54.17%*	*96.34%*	*60.24%*	*44.23%*	*91.78%*	56.01%	52.57%	93.07%
	PUNC ROUGE (1_precision)	*28.76%*	*38.10%*	*98.59%*	*54.77%*	*55.49%*	*96.02%*	*27.66%*	*37.42%*	*98.75%*	*27.39%*	*37.23%*	*98.53%*	34.64%	42.06%	97.97%
	PUNC ROUGE (L_precision)	*27.33%*	*37.87%*	*99.16%*	*54.51%*	*55.87%*	*96.11%*	*27.31%*	*37.92%*	*99.35%*	*27.33%*	*37.83%*	*99.36%*	34.12%	42.37%	98.50%
	PUNC BERT (precision)	***85.96%***	***84.23%***	***73.54%***	***79.47%***	***78.80%***	***87.05%***	*54.78%*	*53.51%*	*94.79%*	***76.04%***	***73.44%***	***84.58%***	**74.06%**	**72.49%**	**84.99%**

Table 8. **Performance comparison all applications (*Vague*, *Adversarial*, *Corrupt Lvl1*, *Corrup Lvl2*)**

T2I Model	Inference Steps	Guidance Scale	Version
SD 1.5	20	7.5	sd-legacy/stable-diffusion-v1-5
SDXL	20	7.5	stabilityai/stable-diffusion-xl-base-1.0
PixArt-Σ	20	4.5	PixArt-alpha/PixArt-Sigma
SDXS	1	None	IDKiro/sdxs-512-0.9

Table 9. Summary of model parameters used in the experiments. All models are sourced from the Hugging Face repository.

LVLM	Instruct	# Params	Version
LLama3.2- Vision	Yes	11B	meta-llama/Llama-3.2-11B-Vision-Instruct
Llava-Next	Yes	7B	llava-hf/llava-v1.6-mistral-7b-hf
Molmo	No	7B	allenai/Molmo-7B-O-0924
QWEN-VL	No	7B	Qwen/Qwen2-VL-7B-Instruct-GPTQ-Int4

Table 10. Summary of model parameters used in the experiments. All models are sourced from the Hugging Face repository.

captions with varying levels of corruption: (1) Level 1: We introduced grammatical mistakes and spelling errors to the prompts from the *Normal* dataset. (2) Level 2: We further intensified corruption by randomly removing 50% of the words in each Level 1 prompt, simulating extreme cases of incomplete or fragmented input. These corrupted prompts are labeled as *Corrupted* and represent varying degrees of aleatoric uncertainty that could interfere with the model's comprehension.

F. Task: Knowledge extraction: Concept

F.1. Deepfake Uncertainty

An intriguing aspect of uncertainty estimation is its potential to benefit other tasks. One pressing issue with diffusion models is the challenge of protecting against harmful applications, such as deepfakes, which pose significant societal risks. As diffusion models continue to improve, the likelihood of misuse grows, raising concerns about the ability to generate realistic images of prominent individuals, such as politicians.

To investigate whether diffusion models have learned to generate images of specific politicians, we focused on the heads of state from the G7 countries. We created 100 prompts with the names of these politicians and used the four previously mentioned diffusion models to generate images based on these prompts. To quantify the uncertainty, we employed a LVLM, but with a modified prompt. Rather than asking the LVLM to describe the image, we directly queried whether a specific politician was present in the image, with a simple "yes" or "no" response. This approach allowed us to assess the LVLM's ability to detect the intended concept—politician recognition—in the generated images. Our findings are summarized in Table 14.

		SDXS			PixArt			SDv1.5			SDXL			Average		
		auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓
microscopic	PUNC ROUGE (LLAVA) (1_recall)	69.23%	67.31%	71.63%	90.36%	92.19%	51.00%	77.52%	75.82%	64.28%	64.80%	58.63%	67.02%	75.48%	73.49%	63.48%
	PUNC ROUGE (LLAVA) (L_recall)	60.80%	61.53%	79.12%	88.80%	89.13%	46.95%	67.08%	68.11%	75.19%	59.14%	55.06%	73.28%	68.95%	68.46%	68.64%
	PUNC BERT (LLAVA) (recall)	69.01%	71.06%	77.56%	92.94%	93.74%	32.58%	74.65%	76.73%	74.55%	61.53%	58.76%	79.43%	74.53%	75.07%	66.03%
	PUNC ROUGE (llama) (1_recall)	30.06%	49.70%	99.82%	58.90%	69.34%	94.50%	43.24%	63.57%	99.94%	52.46%	64.84%	99.14%	46.17%	61.86%	98.35%
	PUNC ROUGE (llama) (L_recall)	29.78%	49.81%	99.69%	58.95%	68.82%	94.00%	46.79%	66.07%	99.59%	52.51%	65.62%	98.99%	47.01%	62.58%	98.07%
	PUNC BERT (llama) (recall)	56.27%	60.44%	87.53%	66.82%	72.59%	83.09%	65.91%	76.15%	82.22%	62.67%	69.48%	86.46%	62.92%	69.66%	84.82%
	PUNC ROUGE (Molmo)(1_recall)	84.00%	85.10%	57.07%	95.47%	97.06%	30.84%	87.24%	90.66%	48.90%	83.56%	84.31%	52.78%	87.57%	89.28%	47.40%
	PUNC ROUGE (Molmo)(L_recall)	87.64%	89.56%	55.28%	96.96%	98.02%	16.85%	**90.20%**	**93.40%**	**43.56%**	**88.19%**	**89.84%**	**47.44%**	**90.75%**	**92.71%**	**40.78%**
	PUNC BERT (Molmo)(recall)	**88.45%**	**89.71%**	**48.68%**	**97.83%**	**98.41%**	**10.32%**	87.75%	91.70%	51.03%	83.10%	85.46%	58.14%	89.28%	91.32%	42.04%
remote sensing	PUNC ROUGE (LLAVA) (1_recall)	58.61%	58.86%	93.11%	92.32%	94.24%	52.38%	50.93%	52.14%	94.84%	44.51%	48.53%	97.63%	61.59%	63.44%	84.49%
	PUNC ROUGE (LLAVA) (L_recall)	58.89%	61.82%	97.65%	94.17%	95.49%	35.64%	53.26%	54.81%	96.51%	54.03%	54.40%	95.88%	65.09%	66.63%	81.42%
	PUNC BERT (LLAVA)(recall)	48.89%	53.08%	95.31%	89.16%	90.98%	57.78%	41.42%	47.63%	97.96%	33.93%	42.77%	98.98%	53.35%	58.62%	87.51%
	PUNC ROUGE (llama) (1_recall)	34.47%	49.21%	99.93%	70.89%	73.16%	84.22%	51.95%	59.10%	99.95%	51.17%	60.32%	99.54%	52.12%	60.45%	95.91%
	PUNC ROUGE (llama) (L_recall)	35.45%	50.23%	99.83%	72.00%	74.78%	80.15%	55.99%	61.70%	99.44%	51.60%	62.11%	99.45%	53.76%	62.21%	94.72%
	PUNC BERT (llama) (recall)	51.96%	57.38%	95.65%	70.80%	73.65%	92.46%	60.68%	60.96%	91.74%	51.64%	56.84%	93.57%	58.77%	62.20%	93.35%
	PUNC ROUGE (Molmo)(1_recall)	85.59%	86.59%	63.87%	94.14%	95.97%	49.29%	73.24%	72.93%	79.56%	73.07%	74.54%	83.48%	81.51%	82.51%	69.05%
	PUNC ROUGE (Molmo)(L_recall)	**89.85%**	**90.63%**	**51.96%**	**97.01%**	**97.94%**	**16.08%**	83.63%	84.65%	**65.35%**	**85.68%**	**87.09%**	**64.00%**	**89.04%**	**90.08%**	**49.34%**
	PUNC BERT (Molmo)(recall)	80.50%	82.56%	75.19%	91.93%	93.08%	43.30%	61.27%	64.74%	90.08%	60.71%	65.32%	90.94%	73.60%	76.42%	74.88%
Texture	PUNC ROUGE (LLAVA) (1_recall)	27.39%	66.97%	99.68%	82.94%	94.90%	84.90%	31.87%	68.29%	91.28%	28.93%	67.95%	99.15%	42.78%	74.53%	93.75%
	PUNC ROUGE (LLAVA) (L_recall)	28.95%	67.52%	99.36%	76.61%	92.26%	87.72%	31.42%	67.96%	91.40%	27.29%	66.86%	99.27%	41.07%	73.65%	94.44%
	PUNC BERT (LLAVA) (recall)	34.28%	71.87%	99.15%	82.56%	94.51%	75.89%	37.21%	72.56%	92.46%	32.34%	70.87%	99.29%	46.60%	77.45%	91.70%
	PUNC ROUGE (llama) (1_recall)	42.02%	75.94%	99.22%	65.48%	87.24%	89.52%	47.66%	80.02%	99.96%	51.95%	80.22%	97.35%	51.78%	80.86%	96.51%
	PUNC ROUGE (llama) (L_recall)	41.84%	76.44%	98.90%	65.09%	87.38%	91.05%	50.45%	81.30%	99.70%	52.43%	81.48%	98.24%	52.45%	81.65%	96.97%
	PUNC BERT (llama) (recall)	**66.38%**	**85.27%**	**80.35%**	75.62%	90.81%	77.67%	70.99%	**87.77%**	**76.84%**	61.85%	**83.97%**	**81.27%**	68.71%	**86.95%**	**79.03%**
	PUNC ROUGE (Molmo)(1_recall)	40.77%	72.88%	95.54%	83.26%	94.80%	73.48%	46.61%	75.64%	**90.87%**	49.16%	78.04%	93.57%	54.95%	80.34%	**88.36%**
	PUNC ROUGE (Molmo)(L_recall)	45.64%	76.90%	95.99%	84.10%	95.13%	74.21%	47.60%	77.70%	93.14%	50.71%	**80.05%**	94.35%	57.01%	82.44%	89.42%
	PUNC BERT (Molmo)(recall)	49.11%	**79.11%**	**94.65%**	**84.40%**	**95.14%**	**69.99%**	47.60%	**78.80%**	93.85%	46.72%	79.20%	96.87%	56.96%	**83.06%**	88.84%

Table 11. **Performance comparison of PUNC applied to different Out of Distribution concepts and different text-to-image models**

		SDXS			PixArt			SDv1.5			SDXL			Average		
		auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓	auroc ↑	aupr ↑	fpr95↓
Vague	PUNC ROUGE (LLAVA) (1_precision)	**99.99%**	**100.00%**	0.00%	**100.00%**	**100.00%**	0.00%	**100.00%**	**100.00%**	0.00%	**99.99%**	**100.00%**	0.00%	**100.00%**	**100.00%**	0.00%
	PUNC ROUGE (LLAVA) (L_precision)	**99.99%**	**100.00%**	0.00%	**100.00%**	**100.00%**	0.00%	**100.00%**	**100.00%**	0.00%	**99.99%**	**100.00%**	0.00%	**100.00%**	**100.00%**	0.00%
	PUNC BERT (LLAVA) (precision)	99.54%	99.95%	1.70%	99.79%	99.98%	0.30%	97.78%	99.75%	9.70%	98.92%	99.92%	5.60%	99.01%	99.90%	4.33%
	PUNC ROUGE (llama)(1_precision)	97.61%	99.58%	12.15%	99.31%	99.89%	2.05%	99.17%	99.85%	3.05%	98.68%	99.81%	4.50%	98.69%	99.78%	5.44%
	PUNC ROUGE (llama)(L_precision)	96.03%	99.35%	19.50%	98.55%	99.77%	6.05%	96.96%	99.51%	13.45%	97.80%	99.67%	10.20%	97.34%	99.58%	12.30%
	PUNC BERT (llama)(precision)	60.63%	92.24%	**82.85%**	67.02%	94.54%	83.10%	37.35%	86.51%	92.70%	70.85%	95.13%	78.00%	58.96%	92.11%	**84.16%**
	PUNC ROUGE (Molmo)(1_precision)	**99.99%**	**100.00%**	0.00%	**100.00%**	**100.00%**	**0.00%**	**99.99%**	**100.00%**	**0.00%**	**99.99%**	**100.00%**	0.00%	**99.99%**	**100.00%**	0.00%
	PUNC ROUGE (Molmo)(L_precision)	**99.99%**	**100.00%**	0.00%	**100.00%**	**100.00%**	**0.00%**	**99.99%**	**100.00%**	**0.00%**	**100.00%**	**100.00%**	0.00%	**100.00%**	**100.00%**	0.00%
	PUNC BERT (Molmo)(precision)	60.63%	92.24%	82.85%	67.02%	94.54%	83.10%	37.35%	86.51%	92.70%	70.85%	95.13%	78.00%	58.96%	92.11%	84.16%
Adversarial	PUNC ROUGE (LLAVA) (1_precision)	*70.83%*	*98.42%*	*47.80%*	*75.17%*	*98.83%*	*48.26%*	*59.03%*	*96.59%*	*49.60%*	*65.03%*	*98.31%*	*43.99%*	67.52%	98.04%	47.41%
	PUNC ROUGE (LLAVA) (L_precision)	*71.61%*	*98.56%*	*49.00%*	*74.33%*	*98.75%*	*48.26%*	59.49%	96.80%	50.00%	64.87%	98.38%	**43.99%**	67.58%	98.12%	47.81%
	PUNC BERT (LLAVA)(precision)	***79.14%***	***99.20%***	***48.00%***	*74.70%*	*98.84%*	*48.06%*	*61.21%*	*97.41%*	*54.20%*	*66.73%*	*98.78%*	*48.11%*	**70.44%**	**98.56%**	**49.59%**
	PUNC ROUGE (llama)(1_precision)	*64.83%*	*98.18%*	*57.03%*	*78.08%*	*99.11%*	*49.20%*	*67.27%*	*98.20%*	*51.50%*	***69.57%***	***98.40%***	*50.60%*	69.94%	98.47%	52.08%
	PUNC ROUGE (llama)(L_precision)	*69.71%*	*98.49%*	*59.64%*	***80.99%***	***99.23%***	***49.00%***	***68.79%***	***98.33%***	*56.11%*	*71.14%*	*98.47%*	*53.40%*	**72.66%**	**98.63%**	**54.54%**
	PUNC BERT (llama)(precision)	*73.98%*	*98.72%*	*78.31%*	*72.95%*	*98.93%*	*86.40%*	*45.09%*	*96.93%*	*94.18%*	*72.69%*	*98.80%*	*81.40%*	66.18%	98.34%	85.07%
	PUNC ROUGE (Molmo)(1_precision)	*64.14%*	*97.77%*	*48.80%*	*76.32%*	*98.90%*	*48.00%*	*63.59%*	*97.73%*	*49.50%*	*62.97%*	*97.64%*	*49.20%*	66.76%	98.01%	48.87%
	PUNC ROUGE (Molmo)(L_precision)	*63.20%*	*97.75%*	*49.00%*	*76.26%*	*98.90%*	*48.40%*	*64.09%*	*97.90%*	***49.70%***	*63.76%*	*97.87%*	*49.60%*	66.83%	98.11%	49.17%
	PUNC BERT (Molmo)(precision)	*73.98%*	*98.72%*	*78.31%*	*72.95%*	*98.93%*	*86.40%*	*45.09%*	*96.93%*	*94.18%*	*72.69%*	*98.80%*	*81.40%*	66.18%	98.34%	85.07%
Corrupt lvl1	PUNC ROUGE (LLAVA) (1_precision)	44.38%	46.02%	96.50%	51.31%	53.87%	96.51%	19.90%	34.21%	99.22%	19.60%	47.23%	99.52%	33.80%	45.33%	97.94%
	PUNC ROUGE (LLAVA) (L_precision)	46.06%	48.93%	97.12%	48.58%	51.26%	96.61%	18.98%	34.00%	99.49%	18.26%	46.95%	99.72%	32.97%	45.29%	98.24%
	PUNC BERT (LLAVA)(precision)	60.71%	65.46%	94.87%	49.73%	52.23%	96.63%	25.07%	36.22%	99.62%	25.65%	50.22%	99.57%	40.29%	51.03%	97.67%
	PUNC ROUGE (llama)(1_precision)	33.10%	43.04%	99.88%	58.67%	62.73%	97.71%	35.83%	42.27%	99.14%	42.49%	46.81%	99.77%	42.52%	48.71%	99.13%
	PUNC ROUGE (llama)(L_precision)	44.92%	48.73%	98.85%	64.78%	65.93%	92.88%	39.64%	43.63%	98.08%	47.27%	48.85%	98.79%	49.15%	51.79%	97.15%
	PUNC BERT (llama)(precision)	**85.85%**	**84.45%**	**73.94%**	**79.43%**	**78.83%**	**87.21%**	**54.97%**	**53.73%**	**94.51%**	**75.76%**	**72.91%**	**84.65%**	**74.00%**	**72.48%**	**85.08%**
	PUNC ROUGE (Molmo)(1_precision)	28.75%	38.34%	98.77%	54.58%	55.29%	96.02%	27.69%	37.57%	98.78%	27.45%	37.27%	98.54%	34.62%	42.12%	98.03%
	PUNC ROUGE (Molmo)(L_precision)	27.29%	38.11%	99.15%	54.32%	55.59%	96.16%	27.30%	38.07%	99.45%	27.35%	37.89%	99.37%	34.07%	42.42%	98.53%
	PUNC BERT (Molmo)(precision)	**85.85%**	**84.45%**	**73.94%**	**79.43%**	**78.83%**	87.21%	**54.97%**	**53.73%**	94.51%	**75.76%**	**72.91%**	84.65%	**74.00%**	**72.48%**	**85.08%**
Corrupt lvl2	PUNC ROUGE (LLAVA) (1_precision)	*25.37%*	*35.74%*	*98.11%*	*62.14%*	*62.43%*	*90.66%*	*26.18%*	*36.05%*	*98.13%*	*27.77%*	*50.14%*	*97.85%*	35.37%	46.09%	96.19%
	PUNC ROUGE (LLAVA) (L_precision)	*25.47%*	*35.87%*	*98.41%*	*60.56%*	*60.22%*	*90.25%*	*25.45%*	*35.94%*	*98.69%*	*26.30%*	*49.90%*	*98.56%*	34.45%	45.48%	96.48%
	PUNC BERT (LLAVA)(precision)	*33.19%*	*39.35%*	*98.28%*	*66.19%*	*66.09%*	*86.84%*	*33.18%*	*39.55%*	*98.58%*	*37.58%*	*55.84%*	*97.01%*	42.54%	50.21%	95.18%
	PUNC ROUGE (llama)(1_precision)	*33.29%*	*42.84%*	*99.90%*	*58.79%*	*62.71%*	*97.92%*	*35.71%*	*42.10%*	*99.14%*	*42.49%*	*46.84%*	*99.78%*	42.57%	48.62%	99.19%
	PUNC ROUGE (llama)(L_precision)	*45.22%*	*48.60%*	*98.82%*	*64.79%*	*65.81%*	*92.98%*	*39.54%*	*43.43%*	*98.18%*	*47.04%*	*48.77%*	*98.76%*	49.15%	51.65%	97.19%
	PUNC BERT (llama)(precision)	***85.96%***	***84.23%***	***73.54%***	***79.47%***	***78.80%***	***87.05%***	*54.78%*	*53.51%*	*94.79%*	***76.04%***	***73.44%***	***84.58%***	**74.06%**	**72.49%**	**84.99%**
	PUNC ROUGE (Molmo)(1_precision)	*28.76%*	*38.10%*	*98.59%*	*54.77%*	*55.49%*	*96.02%*	*27.66%*	*37.42%*	*98.75%*	*27.39%*	*37.23%*	*98.53%*	34.64%	42.06%	97.97%
	PUNC ROUGE (Molmo)(L_precision)	*27.33%*	*37.87%*	*99.16%*	*54.51%*	*55.87%*	*96.11%*	*27.31%*	*37.92%*	*99.35%*	*27.33%*	*37.83%*	*99.36%*	34.12%	42.37%	98.50%
	PUNC BERT (Molmo)(precision)	***85.96%***	***84.23%***	***73.54%***	***79.47%***	***78.80%***	***87.05%***	*54.78%*	*53.51%*	*94.79%*	***76.04%***	***73.44%***	***84.58%***	**74.06%**	**72.49%**	**84.99%**

Table 12. **Performance comparison all applications (*Vague*, *Adversarial*, *Corrupt Lvl1*, *Corrup Lvl2*)**

One key insight from this study is that the effectiveness of concept recognition depends heavily on the LVLM model's ability to identify the targeted concept in an image. To assess Molmo's capacity for concept extraction, we evaluated it on 20 images of the selected politicians from the web, along with images of individuals who were not politicians. The results of Molmo's performance in recognizing each politician are shown in Table 13 are as follows:

Name	Precision (%)	Recall (%)
Joe Biden	83	100
Emmanuel Macron	82	100
King Charles III	100	100
Justin Trudeau	88	100
Kishida Fumio	83	33
Giorgia Meloni	61	73
Olaf Scholz	55	100

Table 13. Performance of Molmo in *Politician* Recognition using Precision and Recall Measures

These results demonstrate varying levels of effectiveness across different politicians. For example, Molmo performs particularly well in identifying King Charles III, achieving perfect precision and recall. In contrast, it struggles more with Giorgia Meloni and Olaf Scholz, where precision and recall scores are lower.

In analyzing the diffusion models, we observed that SDXL was the most consistent in generating recognizable images of Emmanuel Macron and demonstrated the highest overall accuracy in generating images for the given prompts. Visual inspection confirmed that SDXL produced the best representations for most prompts, showcasing its relative reliability in political figure generation.

F.2. Copyright Uncertainty

Another important category of concepts worth evaluating is related to copyright. Specifically, we explore whether a diffusion model can generate well-known cartoon characters from specific brands, such as Mickey Mouse, Donald Duck, Darth Vader, and Pikachu. This experiment enables us to assess the extent to which the model has learned to generate these copyrighted concepts.

Following a similar approach to the previous experiments, we generated 100 prompts for each character concept and used the various diffusion models to generate corresponding images. We then employed the LVLM component of PUNC to determine whether the target character was present in each generated image, asking the LVLM model to provide a simple ”yes” or ”no” answer regarding the presence of each concept.

The results of this experiment are presented in Table 15. Notably, while PixArt performed less effectively in the earlier experiments with politicians, it emerges as one of the top performers in this context, successfully generating recognizable representations of copyrighted characters. This contrast suggests differences in the training datasets used for these diffusion models and highlights how these variations can impact model behavior in response to regulatory constraints and copyright considerations.

F.3. Task: Bias of Diffusion Model

Previous work [3] has shown that diffusion models may exhibit gender and racial biases when generating people from vague prompts. In our experiments, we explicitly instructed diffusion models to generate individuals performing one of 13 specific jobs, including ‘CEO’, ‘basketball player’, ‘call center employee’, ‘cleaning staff’, ‘computer user’, ‘firefighter’, ‘marketing professional’, ‘medical doctor’, ‘nurse’, ‘police officer’, ‘politician’, ‘rap singer’, and ‘teacher’.

We specified both the gender and race of the person to be generated, limiting the racial categories to white, Black, and Asian to simplify the analysis, as results became inconsistent with a larger variety of races. Similar to previous experiments, we used a LVLM to confirm whether the specified concepts were accurately generated using a simple yes/no question.

The results of this study, shown in Figures 6 and 7, reveal significant fairness issues in SDv1.5, which tend to diminish in newer models like PixArt. Interestingly, however, PixArt still exhibits biases; for example, it struggles to generate a white basketball player or a white rap singer, while defaulting to certain racial stereotypes for these jobs. These findings suggest that understanding and addressing uncertainty in diffusion models could be an important step toward mitigating such biases.

G. Qualitative Results

In this section, we present qualitative results from the images generated and the captions generated from our studies. In Fig. 8 Fig. 9 Fig. 10 Fig. 11 , we present generated captions of microscopic images. In Fig. 12 Fig. 13 Fig. 14 Fig. 15 , we present generated captions of Remote Sensing images. In Fig. 16 Fig. 17 Fig. 18 Fig. 19 , we present generated captions of Texture images. The Fig. 20 presents examples of images generated using the same prompt in different T2I models. Fig. 21 presents examples of **Microscopic** images generated using the same prompt in different T2I models. Fig. 22 presents examples of **Texture** images generated using the same prompt in different T2I models. Fig. 23 presents examples of **Remote Sensing** images generated using the same prompt in different T2I models.

	Justin Trudeau	King Charles	Fumio Kishida	Olaf Scholz	Emmanuel Macron	Joe Bidden	Giorgia Meloni	AVERAGE
SDXL	94	94	88	88	98	96	21	**95.5**
SDv1.5	78	60	62	88	94	80	18	78
SDXS	80	40	70	82	94	76	28	72.5
PixArt	36	76	62	78	94	90	15	74
AVERAGE	72	67.5	70.5	84	**95**	85.5	20.5	70.71

Table 14. **Performance comparison *Politicians* Application**: Accuracy (%) Measures

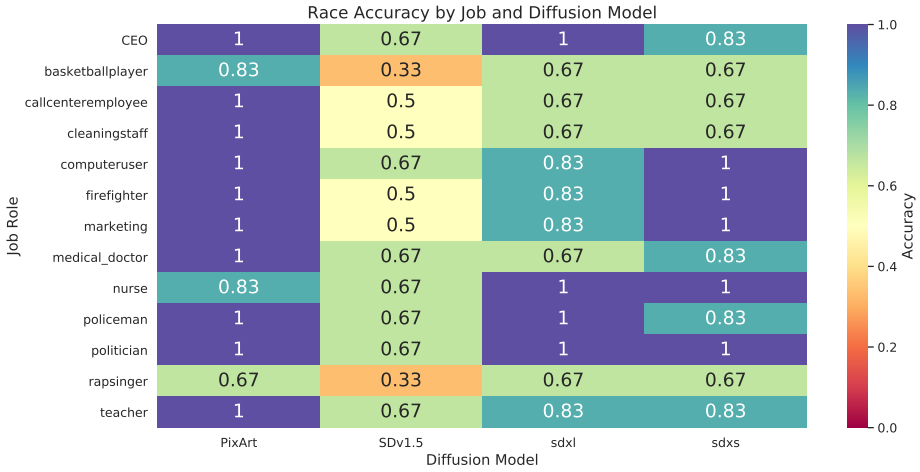


Figure 6. Illustration of gender bias in diffusion models with respect to job representation.

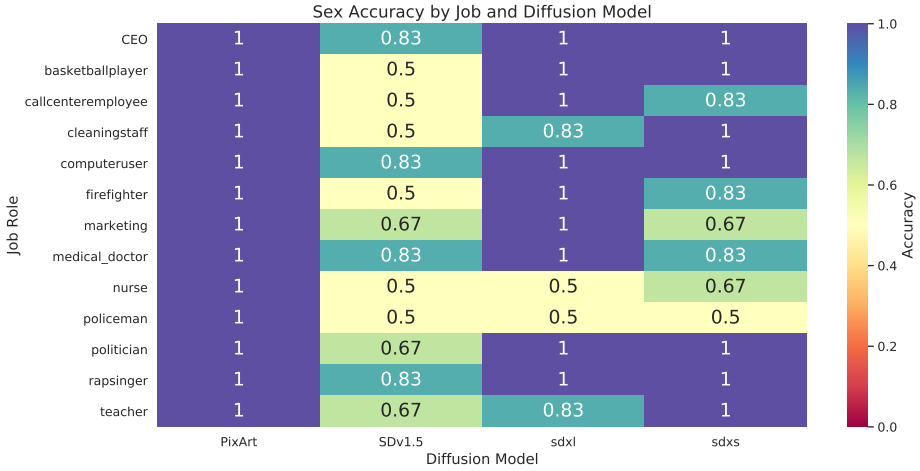


Figure 7. Illustration of racial bias in diffusion models with respect to job representation.

	Darth Vader	Pikachu	Mickey	Donald	Average
SDv1.5	96	88	86	8	87.5
PixArt	96	100	100	94	97.5
SDXS	100	96	96	90	95.5
SDXL	98	100	94	96	97
AVERAGE	97.5	96	94	90	94.3

Table 15. Performance of Molmo in *Copyrights* Recognition: Accuracy (%) Metrics

Figure 8. Example of a generated microscopic image with PixArt-Σ and the caption generated

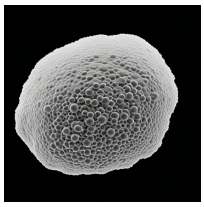
	original prompt	This is a microscopic image. The image you've provided appears to be a microscopic view of a substance, possibly a biological sample given the presence of cells. The cells are spherical and have a darker center, which could indicate the nucleus, and a lighter periphery, which might be the cell membrane. The cells are densely packed together, suggesting a high cell density.
	llava caption	The image appears to be a close-up of a spherical object with a textured surface. The texture looks like a series of small, rounded protrusions or indentations, giving it a somewhat porous or spongy appearance. The background is completely black, which contrasts with the white or light-colored surface of the object, making it stand out. The image is monochromatic, with no color information, and it has a high-resolution quality, suggesting it might be a scientific or technical photograph, possibly of a material or a natural object with a complex surface structure.
	llama caption	This image presents a detailed scanning electron microscope image of a pollen grain, featuring a round shape and a textured surface. The grain is composed of numerous small circles, resembling bubbles, and is centered on a black background. The image appears to be a zoomed-in view of a pollen grain, possibly for the purpose of identification.
	molmo caption	The photo shows a fascinating spherical structure with a complex internal pattern. The outer surface is white and textured, resembling a spongy or foam-like material. Inside, there's an intricate arrangement of smaller spheres or bubbles, creating a honeycomb-like structure. This internal pattern is particularly interesting, as it appears to be in a state of constant motion. The smaller spheres are shifting and rearranging themselves, giving the impression of a dynamic, fluid-like environment within the larger structure. The overall appearance is reminiscent of a microscopic organism or a highly magnified view of a complex material. The contrast between the white outer surface and the darker inner pattern is striking, making the structure stand out vividly against the black background. This image captures a moment of intricate beauty and complexity, showcasing the fascinating structures that can exist at microscopic levels. The motion of the inner spheres adds an extra layer of visual interest, making the viewer feel as if they're witnessing a living, ever-changing entity.

Figure 9. Example of a generated microscopic image with SDXL and the caption generated

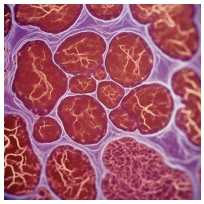
	original prompt	This is a microscopic image. The image appears to be a microscopic view of a tissue sample, possibly from a histological slide prepared for examination under a microscope. The tissue is composed of various cell types and extracellular matrix, which are typical components of tissue samples.
	llava caption	The image appears to be a microscopic view of a tissue sample, possibly from a histological slide prepared for examination under a microscope. The tissue is stained with a purple-red stain, which is commonly used in histology to differentiate between various cell types and tissue components. The staining highlights the cellular structures and extracellular matrix, which can provide information about the tissue's health and function. The specific details of the tissue, such as the presence of cells, glands, or other structures, would require a more detailed examination by a trained professional.
	llama caption	The image presents a detailed microscopic view of muscle tissue, showcasing its intricate structure and composition. **Key Features:** * **Muscle Fibers:** The image features numerous muscle fibers, each surrounded by a distinct border. * **Fiber Structure:** Each fiber is encased in a thick, purple membrane, with a lighter purple inner layer and a darker outer layer. * **Inner Layer:** The inner layer is characterized by a network of thin, yellow lines that resemble lightning bolts or tree branches, adding to the visual complexity of the image. * **Background:** The background of the image is a light purple color, providing a subtle contrast to the darker tones of the muscle fibers. **Overall Impression:** The image presents a highly detailed and visually striking representation of muscle tissue, with its intricate structure and composition making it a fascinating subject of study.
	molmo caption	The photo shows a collection of red, oval-shaped objects with intricate white patterns inside them. These objects are arranged in a pattern and surrounded by a light purple background. The white patterns within the red objects resemble veins or neural networks, creating a complex and visually striking image. The overall composition gives the impression of a microscopic view of biological structures, possibly cells or tissue samples. The contrast between the red objects, white patterns, and light purple background creates a visually appealing and intriguing image.

Figure 10. Example of a generated microscopic image with SDXS and the caption generated

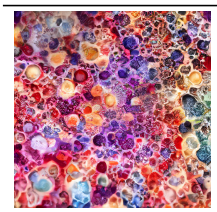
	original prompt	This is a microscopic image. The image appears to be a microscopic view of a tissue sample, possibly from a histological slide prepared for examination under a microscope. The tissue is composed of various cell types and extracellular matrix, which are typical components of tissue samples.
	llava caption	The image appears to be an abstract or surreal photograph. It features a dense, colorful pattern that resembles a close-up of a textured surface or possibly a microscopic view of cells or particles. The colors are vibrant and varied, with no clear, recognizable shapes or forms. The overall effect is one of a chaotic, yet harmonious, composition that invites the viewer to interpret the image in their own way.
	llama caption	This image showcases a vibrant, abstract composition featuring a kaleidoscope of colors and shapes. The background is a swirling mix of various hues, including pink, purple, orange, blue, red, and black, with some areas appearing more opaque than others. The colors are blended and layered, creating a sense of depth and texture. The overall effect is one of dynamic movement and energy, as if the colors are swirling and dancing across the surface.
	molmo caption	The photo shows a collection of colorful rocks or crystals. These minerals display a variety of vibrant hues including red, purple, blue, green, yellow, and white. The rocks appear to be of different sizes and shapes, creating an interesting and diverse arrangement. The minerals are tightly clustered together, forming a dense pile or stack. This close grouping of various colored stones creates a visually striking composition, with the different colors and textures of the rocks contrasting against one another. The image focuses solely on these mineral specimens, showcasing their natural beauty and diversity without any additional elements or background distractions. The arrangement highlights the unique characteristics of each individual rock while also demonstrating how they come together to form a visually appealing collection.

Figure 11. Example of a generated microscopic image with SDv1.5 and the caption generated

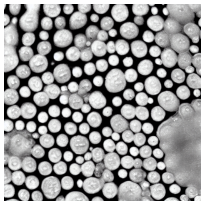
	original prompt	This is a microscopic image. The image appears to be a microscopic view of a tissue sample, possibly from a histological slide prepared for examination under a microscope. The tissue is composed of various cell types and extracellular matrix, which are typical components of tissue samples.
	llava caption	The image appears to be a microscopic view of a substance, possibly a biological tissue or a material with a porous structure. The contents are composed of numerous spherical particles or cells, which are densely packed together. The image is in black and white, which suggests that it might be a scientific or medical photograph, used to study the microstructure of the material or tissue.
	llama caption	The image depicts a collection of variously sized circles, all of which are gray and white, with some having a textured surface. They are arranged in a pattern that resembles a honeycomb, with some of them appearing to be connected to each other. The background is black, which makes the circles stand out.
	molmo caption	The photo shows a collection of small, round objects that resemble pebbles or stones. These objects vary in size and shape, creating an interesting texture and pattern. The image is in black and white, which emphasizes the contrast between the light-colored objects and the dark background. The objects appear to be arranged in a somewhat random pattern, giving the impression of a natural collection or perhaps a close-up view of a rocky surface. Some of the objects have slightly irregular shapes or small indentations, adding to their natural appearance. The black and white nature of the image accentuates the textures and variations in the objects, making them stand out more prominently against the dark background. This monochromatic approach creates a striking visual effect, highlighting the forms and subtle differences between each individual object. Overall, the photo presents an intriguing composition of small, round forms, showcasing the natural variations and textures found in a collection of similar objects.

Figure 12. Example of a generated Remote Sensing image with PixArt-Σ and the caption generated

	original prompt	This is a remote sensing image. The image appears to be an aerial or satellite photograph of a landscape, possibly a rural or agricultural area. The image is divided into several distinct sections, each with its own characteristics.
	llava caption	The photo appears to be an aerial or satellite image of a landscape that includes agricultural fields. The fields are divided into sections, possibly for crop rotation or different types of crops. There are green areas that could be vegetation, and some areas that might be fallow or not currently being used for agriculture. The roads and pathways suggest that this is a rural area with some level of infrastructure. The image is quite detailed, allowing for the identification of different types of crops and the layout of the fields.
	llama caption	The image is an aerial view of a network of roads and fields. The image is a bird's eye view of a road network that crosses a series of fields. The roads are mostly straight, but they intersect with each other, creating intersections and turns. The roads are mostly white, with some sections of road being darker than others, possibly due to the color of the ground they are built on. The fields are a mix of green and brown colors, with some areas being more vibrant than others. The fields are rectangular in shape, with some being larger than others. There are also some smaller, more irregularly shaped fields scattered throughout the area. In the background, there are some trees and other vegetation visible, adding depth and texture to the image. The overall effect is one of a complex network of roads and fields, with a mix of different colors and textures.
	molmo caption	This photo presents an aerial view of a diverse landscape divided into distinct sections. The image showcases a patchwork of different land uses, including areas of green vegetation, likely trees or crops, and brown sections that appear to be plowed fields. A network of white roads crisscrosses the landscape, connecting various parts of the land. The layout suggests a planned development, with the roads forming a grid-like pattern that divides the land into rectangular plots. This arrangement is typical of agricultural or urban planning practices, where different areas are allocated for specific purposes. The contrast between the green and brown sections is particularly striking. The green areas likely represent areas of active vegetation or crop growth, while the brown sections could be fallow fields or areas prepared for future cultivation. Overall, this aerial view provides a comprehensive look at how different types of land are organized and managed within a larger area, highlighting the interplay between natural landscapes and human infrastructure.

Figure 13. Example of a generated Remote Sensing image with SDXL and the caption generated

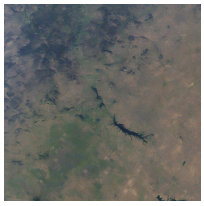
	original prompt	This is a remote sensing image. The image you've provided appears to be a remote sensing image, likely from a satellite or aerial sensor. It shows a dark, textured pattern that could represent a variety of features depending on the context. The image is highly pixelated and lacks clear, distinct features that would allow for a detailed analysis.
	llava caption	The photo appears to be an aerial or satellite image of a landscape. It shows a pattern of fields or plots, with some areas that might be water bodies or wetlands. There are also some darker areas that could be forests or areas with dense vegetation. The overall color palette is muted, with earth tones dominating the image. The image is somewhat blurred, which makes it difficult to discern specific details.
	llama caption	This image presents a satellite photograph of a landscape, showcasing a vast expanse of land with numerous small patches of greenery, interspersed with areas of brown and gray. The land is dotted with a network of rivers and streams, featuring a prominent river running diagonally across the center of the image. **Key Features:** * **Land Cover:** The landscape is predominantly covered in brown and gray hues, with smaller areas of green. * **Water Features:** Several rivers and streams are visible, including a prominent river that runs diagonally across the center of the image. * **Geographic Context:** The image appears to be a satellite photograph of a rural area, possibly in a semi-arid or arid region, given the brown and gray tones. * **Resolution:** The image has a high resolution, allowing for clear visibility of the land features and water bodies. **Visual Hierarchy:** * The river is the most prominent feature in the image, running diagonally across the center.
	molmo caption	The photo shows a satellite view of a diverse landscape. The terrain is predominantly brown, with various shades creating a textured appearance. In the upper left corner, there's a darker area that could be a forest or dense vegetation. The lower left portion features patches of green, likely representing grasslands or fields. A notable feature is a dark, irregular shape in the center, which appears to be a river or stream. The image captures the natural variation of the Earth's surface from a high altitude, providing a bird's-eye view of the land's features and topography.

Figure 14. Example of a generated Remote Sensing image with SDXS and the caption generated

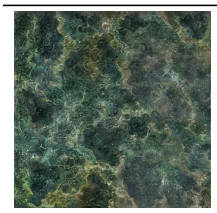
	original prompt	This is a remote sensing image. The image you've provided appears to be a remote sensing image, likely from a satellite or aerial sensor. It shows a dark, textured pattern that could represent a variety of features depending on the context. The image is highly pixelated and lacks clear, distinct features that would allow for a detailed analysis.
	llava caption	The image appears to be a close-up of a natural landscape, possibly a forest or a dense vegetation area. The photo is taken from a low angle, which gives a sense of being immersed in the greenery. The colors are rich and varied, with shades of green and hints of other colors that could be moss, lichen, or other plant life. The texture is quite detailed, suggesting a high-resolution image that captures the intricacies of the environment. The overall effect is one of depth and complexity, with the viewer's eye drawn into the dense foliage.
	llama caption	The image depicts a highly detailed aerial view of a rural area with a diverse landscape, featuring numerous small villages, roads, and fields. The image is composed of a vast array of colors, including dark green, light green, tan, brown, and gray, which create a natural and earthy tone. The landscape is densely populated with small villages and settlements, which are interspersed with numerous winding roads, fields, and forested areas. The overall effect of the image is one of a lush, rural environment with a mix of natural and human-made features, creating a visually striking and detailed representation of the area.
	molmo caption	The photo shows a close-up view of a textured surface, likely a rock or stone. It has a complex, mottled appearance with various shades of green, gray, and brown. The surface is uneven and bumpy, with small crevices and holes visible throughout. This intricate pattern creates an interesting visual texture, resembling natural geological formations.

Figure 15. Example of a generated Remote Sensing image with SD1.5 and the caption generated

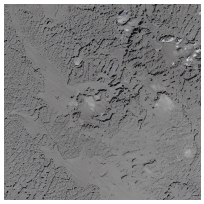
	original prompt	This is a remote sensing image. The image you've provided appears to be a remote sensing image, likely from a satellite or aerial sensor. It shows a dark, textured pattern that could represent a variety of features depending on the context. The image is highly pixelated and lacks clear, distinct features that would allow for a detailed analysis.
	llava caption	The image appears to be a close-up of a textured surface with a pattern that could be interpreted as a natural or geological formation. The texture is rough and uneven, with darker areas that might suggest shadows or crevices. The image is somewhat abstract and lacks clear, identifiable features, making it difficult to determine the exact nature of the surface. It could be a rocky surface, a close-up of a fabric with a rough texture, or even a close-up of a surface with a pattern that resembles a natural or geological formation.
	llama caption	This image depicts a gray surface with a rough, textured appearance, featuring a variety of small ridges, grooves, and depressions. The surface is characterized by a rough, rocky texture, with numerous small depressions and irregularities that give it a rugged appearance. The overall texture of the surface is rough and uneven, with many small features that give it a complex and varied appearance.
	molmo caption	The photo shows a textured surface with a rough, uneven appearance. It has a predominantly gray color with variations in shade, creating a mottled effect. The surface features numerous small indentations and ridges, giving it a complex topography. In the center of the image, there's a notable concentration of darker gray areas, which stand out against the lighter surrounding regions. This central area appears to have more pronounced textures and variations. The overall impression is of a close-up view of a rough material, possibly a rock, concrete, or a textured wall. The intricate patterns and uneven surface create an interesting visual texture, showcasing the natural variations and imperfections of the material. The image is square-shaped and focuses entirely on this textured surface, without any other objects or elements present.

Figure 16. Example of a generated Texture image with PixArt-Σ and the caption generated

	original prompt	This is a texture image. The image displays a close-up view of a textured surface that resembles a wall or a piece of fabric. The texture is characterized by a series of interconnected lines and cracks that create a pattern across the entire visible area. The lines are irregular and vary in width, suggesting a natural or organic origin for the texture. The color palette is limited, with the dominant hue being a muted gray, which gives the surface a somewhat aged or weathered appearance.
	llava caption	The photo shows a close-up of a textured surface that resembles a cracked or fragmented pattern. The texture appears to be a series of interconnected lines or cracks, creating a complex, geometric design. The color palette is quite muted, with shades of gray and white dominating the image. The surface could be a material with a natural or man-made pattern, such as a stone, concrete, or a textured paint finish. The image does not contain any recognizable objects or symbols that would provide a clear context or identification of the material.
	llama caption	The image depicts a gray surface with a cracked pattern, resembling the dry, cracked earth. The surface is light gray with brown lines that resemble cracks, and it fills the entire image.
	molmo caption	The photo shows a close-up view of a cracked, dry surface. The cracks form an intricate pattern across the entire image, creating a complex network of lines. The surface appears to be a light gray or white color, with the cracks appearing darker in contrast. This cracked texture is reminiscent of dried skin, parched earth, or perhaps a weathered wall. The detailed pattern of the cracks gives the image a sense of texture and depth, making it visually interesting despite its simple composition. The overall effect is one of aridity and age, evoking thoughts of drought-stricken landscapes or the passage of time on man-made structures. The intricate nature of the cracks adds a layer of complexity to an otherwise stark and barren-looking surface.

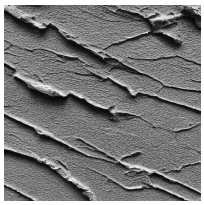
Figure 17. Example of a generated Texture image with SDXL and the caption generated

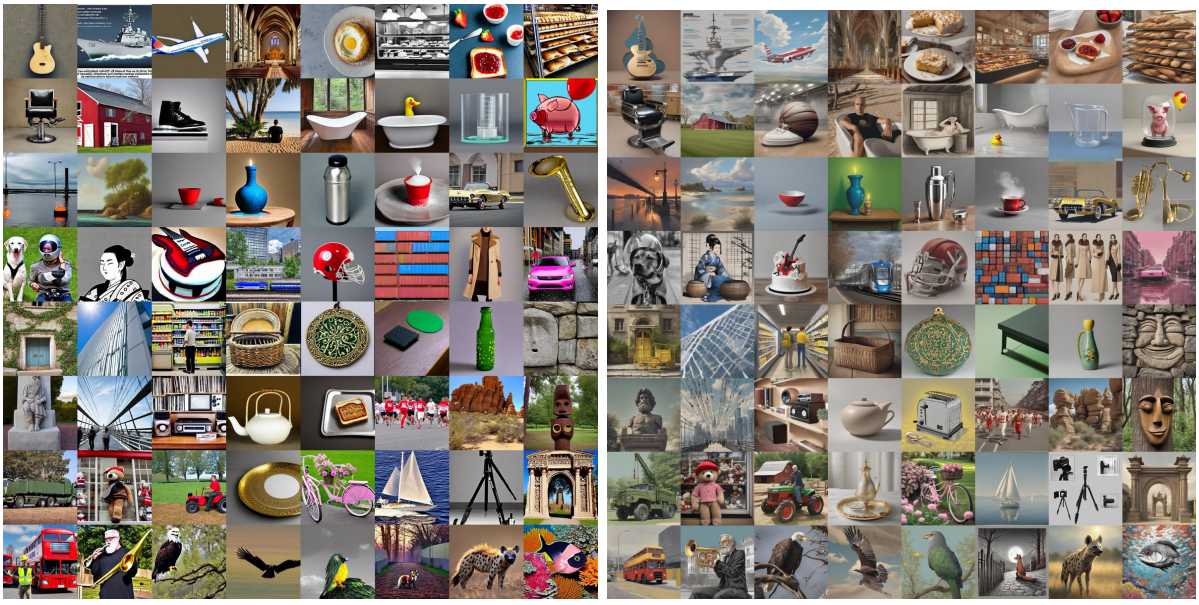
	original prompt	This is a texture image. The photo shows a close-up of a surface with a pattern of blue and green circles. The texture appears to be smooth with a glossy finish, as indicated by the reflection of light on the surface. The circles are uniform in size and are arranged in a seemingly random pattern.
	llava caption	The photo appears to be a close-up of a patterned surface, possibly a fabric or a wallpaper with a repeating design. The design consists of a series of circles in various shades of green, blue, and teal, creating a mosaic-like effect. The circles are arranged in a way that gives the impression of depth and texture, with some circles appearing larger and closer to the viewer, while others are smaller and further away. The overall effect is one of a vibrant, organic pattern that could be used for decorative purposes.
	llama caption	The image presents a visually striking composition, featuring a multitude of small, circular shapes in various shades of green and blue, arranged in a seemingly random pattern. The circles are densely packed, with no discernible gaps or spaces between them. **Color Palette:** * The dominant colors are shades of green, ranging from light to dark, with occasional blue accents. * The green tones are predominantly light and medium, with a few darker shades scattered throughout. * The blue tones are less prevalent, but add a pop of contrast to the overall design. **Circle Characteristics:** * The circles are uniform in size and shape, with no visible irregularities or distortions. * They are densely packed, with no apparent gaps or spaces between them. * The edges of the circles are sharp and defined, with no blurring or softening. **Background:** * The background is a light gray or beige color, which provides a subtle contrast to the vibrant colors of the circles. **Overall Impression:
	molmo caption	The photo shows a collection of small, circular objects arranged in a dense pattern. These circles vary in color, including shades of green, blue, white, and black. They appear to be made of glass or plastic and have a smooth, glossy surface. The objects are tightly packed together, creating a mosaic-like effect. The overall composition is square-shaped, with the circles filling the entire frame. The arrangement gives a sense of order and symmetry, while the variety of colors adds visual interest and depth to the image.

Figure 18. Example of a generated Texture image with SDXS and the caption generated

	original prompt	This is a texture image. The photo shows a close-up of a textured surface that appears to be a wall or a piece of fabric with a rough, cracked pattern. The texture is characterized by a series of interconnected lines and cracks that give it a cracked or peeling appearance. The color palette is quite muted, with shades of gray and white
	llava caption	The photo shows a close-up of a textured surface that appears to be cracked or broken. The texture is rough and uneven, with a pattern of cracks and lines that suggest a material that has been subjected to stress or has dried out. The colors are muted, with a range of grays and whites, which gives the image a somewhat monochromatic look. The exact nature of the material is not clear from the image alone, but it could be concrete, stone, or another similar material.
	llama caption	The image is a close-up photograph of what appears to be a cracked and weathered surface, possibly a wall or a piece of wood. The surface is a grayish color and features numerous cracks, some of which are quite deep and jagged. The cracks are not uniform in size or shape, with some being wider and more prominent than others. The surface appears to be dry and cracked, with a rough texture that suggests it may be made of a material like wood or stone. The cracks give the surface a worn and weathered appearance, as if it has been exposed to the elements for an extended period. Overall, the image suggests a sense of age and wear, with the cracked and weathered surface conveying a sense of history and deterioration.
	molmo caption	The photo shows a close-up view of a dry, cracked surface. The texture is predominantly gray with various shades, ranging from light to dark. The surface is covered in a complex network of cracks, creating an intricate pattern across the entire image. The cracks vary in thickness and direction, giving the surface a rugged and weathered appearance. Some cracks are thin and delicate, while others are thicker and more prominent. The overall effect is reminiscent of parched earth or dried mud, with the cracks creating a web-like structure across the gray background. The image is entirely focused on this cracked surface, with no other objects or elements visible. The lighting appears to be even, highlighting the texture and details of the cracks without creating harsh shadows. This close-up view emphasizes the natural patterns and textures that can occur in dried surfaces, showcasing the intricate and sometimes beautiful results of natural processes like drying or weathering.

Figure 19. Example of a generated Texture image with SD1.5 and the caption generated

	original prompt	This is a texture image. The photo shows a close-up of a textured surface that appears to be a wall or a piece of fabric with a rough, cracked pattern. The texture is characterized by a series of interconnected lines and cracks that give it a cracked or peeling appearance. The color palette is quite muted, with shades of gray and white
	llava caption	The photo shows a close-up of a textured surface with a pattern that resembles cracks or lines. The texture appears to be a combination of smooth and rough surfaces, possibly suggesting a material with a rough grain or a surface with a patterned to look like a natural or man-made texture. The colors are muted, with a grayish tone dominating the image. The image does not contain any recognizable objects or symbols that would provide a clear context or subject matter.
	llama caption	The image depicts a grey, textured surface with a rough, cracked texture, resembling a wall covered in thick, uneven plaster. The surface is composed of several layers of thick, uneven plaster, with each layer slightly overlapping the next, creating a rough and irregular texture. The color is a uniform grey, with no visible patterns or designs. The overall effect is one of ruggedness and imperfection, as if the surface has been weathered and worn over time.
	molmo caption	The photo shows a close-up view of a textured surface with a complex pattern of cracks and crevices. The surface appears to be gray in color and has a rough, uneven texture. The cracks are predominantly white and run diagonally across the image, creating a network of fissures that give the surface a weathered or aged appearance. The overall effect is reminiscent of dried, cracked earth or possibly a textured wall or ceiling. The intricate pattern of cracks and the interplay of light and shadow on the surface create an interesting visual texture that draws the eye across the entire frame.



Stable Diffusion 1.5 SDXL

SDXS PixArt Σ

Figure 20. **Normal** images with different models. The same prompts are used to generate the images with the four models

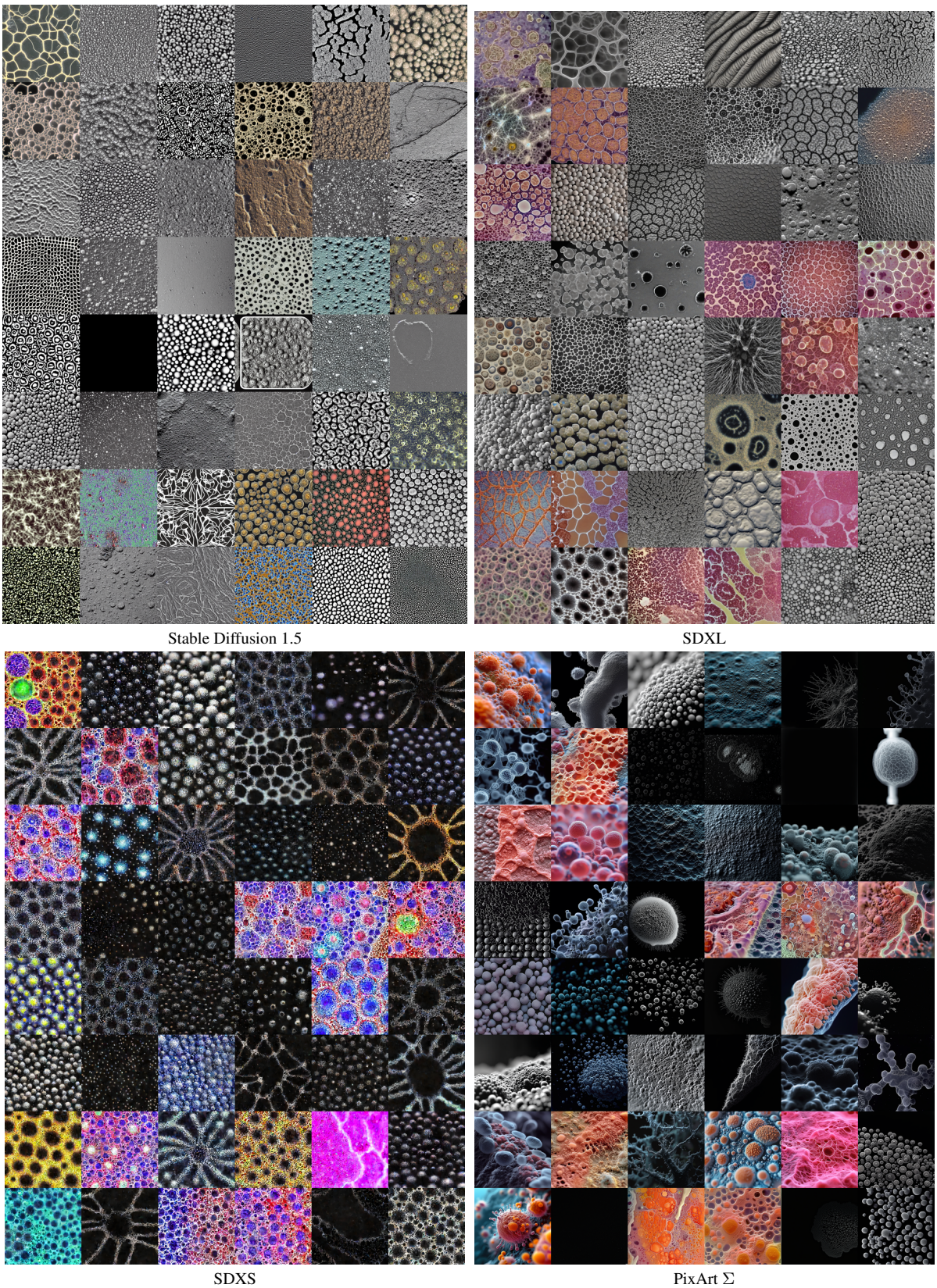
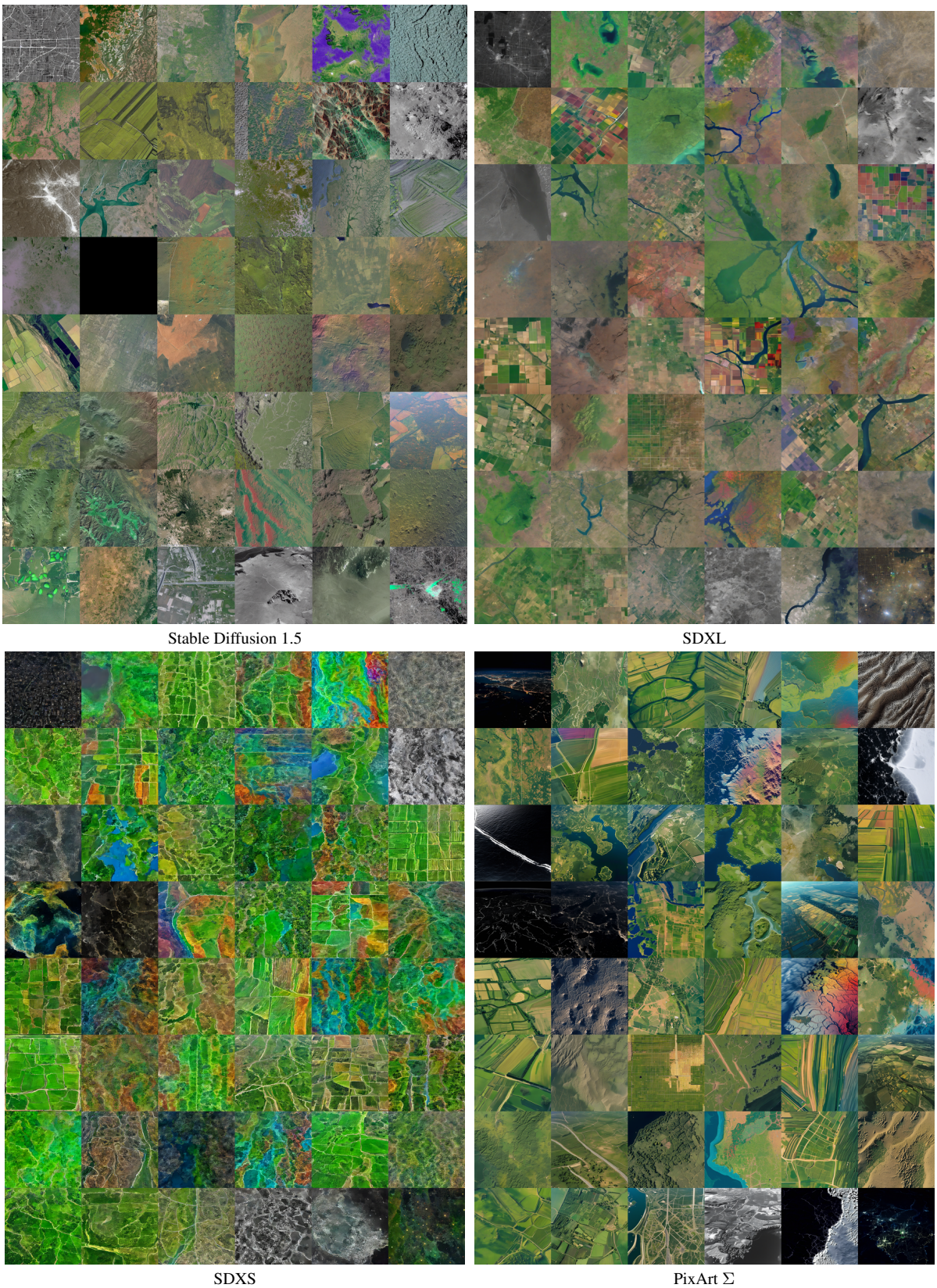


Figure 21. **Microscopic** images with different models. The same prompts are used to generate the images with the four models

Stable Diffusion 1.5 SDXL

SDXS PixArt Σ

Figure 22. **Texture** images with different models. The same prompts are used to generate the images with the four models



Stable Diffusion 1.5 | SDXL

SDXS | PixArt Σ

Figure 23. **Remote Sensing** images with different models. The same prompts are used to generate the images with the four models