**I. Image Prompt Generation**

For all the three datasets ‘MS COCO’, ‘Visual Genome’, and ‘CC3M’, the selected original caption file is split into 16 small caption files, each has 5000 selected original image captions, so that the image caption generation tasks can be executed in parallel.

We will totally generate 3 different types of image prompts, i.e.,

* image prompts by concatenating the original image captions with 5 prompt templates (raw\_prompt):

1. "a high fidelity, high resolution image of "
2. "a high fidelity, high resolution picture of "
3. "a high fidelity, high resolution photo of "
4. "a high fidelity, high resolution photograph of "
5. "a high fidelity, high resolution, realistic image of "

* Image prompts by paraphraser re-generated captions with Meta-Llama-3-8B-Instruct (para\_prompt).
* Image prompts generated by image captioner, in our case, we use LLaVA\_NeXT (cap\_prompt).

For each type of image prompts, for each selected image, we totally generate 10 prompts for it. We totally selected 80000 images from each of the 3 datasets.

**The original caption characteristic for the 3 datasets:**

* MS COCO – human annotated captions
* CC3M – concept captions, i.e., image alternative text with further multi-round concept filtering
* Visual Genome – dense captions, 5.4 Million Region Descriptions, 3.8 Million Object Instances, 2.8 Million Attributes, 2.3 Million Relationships

**Experiments**

Currently we finished only part of the prompt generation experiments. (e.g., for MS COCO dataset, we finished the three types of image prompts, i.e., raw\_prompt, para\_prompt, cap\_prompt, from sel\_caption\_file 5000 to sel\_caption\_file 15000. We still need to generate the prompts for sel\_caption\_file {20000, 25000, …, 80000}

For CC3M, Visual Genome we need to do the same. But we need to prepare the sel\_caption\_file at first.

**II. Video Prompt Generation**

For video caption generation, we adopt 2 video datasets, MSR-VTT and CelebV-Text.

**The original caption characteristic for the 2 datasets:**

* MSR-VTT – human generated video captions
* CelebV-Text – dense facial video captions with 40 general appearances, 5 detailed appearances, 6 light conditions, 37 actions, 8 emotions, and 6 light directions

**Experiments:**

Following the same configurations as image prompt generation, for each video, we generate 3 different types of video prompts: raw\_prompt, para\_prompt, and cap\_prompt. We totally selected 10000 videos from each of the 2 video datasets.

**III. Image Generation**

Based on the generated 3 types of image prompts, we adopt 5 image generation models to generate the text prompts related images.

The adopted 5 image generation models are as below:

1. DeepFloydIF
2. Kandinsky3
3. PixArt\_Σ
4. StableDiffusion3
5. StableDiffusionXL

**Experiments:**

T2I; I2I; different scales; different random seeds;

For each of 80000 selected original text-image pair, using the generated 3 types of text prompts to generate images with 5 image generation models.

**IV. Video Generation**

Based on the generated 3 types of video prompts, we adopt 5 Video generation models to generate the text prompts related videos.

1. GEN-3 Alpha
2. Open-Sora
3. VideoCrafter2
4. LaVie-2
5. VEnhancer

**Experiments:**

T2V; different scales; different random seeds;

For each of 10000 selected original text-video pair, using the generated 3 types of text prompts to generate videos with 5 image generation models.

**V. Image/Video Dataset Profiling**

**Image Profiling**

1. DCT
2. DFT
3. DFT Power Spectrum
4. GLCM
5. Image Texture Descriptors, i.e., LBP, CoALBPs, LPQ

**Video Profiling**

5 image profilers plus 2 video specific profilers

1. EvalCrafter
2. VBench

**VI. Deep Fake Detection**

35 SOTA detectors:

* 5 Naive Detectors: Xception, MesoNet, MesoInception, CNN-Aug, EfficientNet-B4
* 19 Spatial Detectors: Capsule, DSP-FWA, Face X-ray, FFD, CORE, RECCE, UCF, Local-relation, IID, RFM, SIA, SLADD, UIA-ViT, CLIP, SBI, PCL-I2G, Multi-Attention, LSDA
* 3 Frequency Detectors: F3Net, SPSL, SRM
* 7 Video Detectors: TALL, I3D, STIL, FTCN, X-CLIP, TimeTransformer, VideoMAE, StA

**Experiments:**

Train all the 35 detectors on all the generated image/video datasets