

## Dataset & Evaluation Metrics

Year	Dataset	Public	Details				
			Category	Images (Resolution)	Annotations	Attrs	Other Information
2008	Oxford-102 Flowers	✓	Flower	8,189 (-)	10	-	-
2011	CUB-200-2011	✓	Bird	11,788 (-)	10	-	BBox, Segmentation...
2014	MS-COCO2014	✓	Iconic Objects	120k (-)	5	-	BBox, Segmentation...
2018	Face2Text	✓	Face	10,177 (-)	1~	-	-
2019	SCU-Text2face	⊕	Face	1,000 (256×256)	5	-	-
2020	Multi-Modal CelebA-HQ	✓	Face	30,000 (512×512)	10	38	Masks, Sketches
2021	FFHQ-Text	✓	Face	760 (1024×1024)	9	162	BBox
2021	M2C-Fashion	⊕	Clothing	10,855,753 (256×256)	1	-	-
2021	CelebA-Dialog	✓	Face	202,599 (178×218)	~5	5	Identity Label...
2021	Faces a la Carte	⊕	Face	202,599 (178×218)	~10	40	-
2021	LAION-400M	✓	Random Crawled	400M (-)	1	-	KNN Index...
2022	Bento800	✓	Food	800 (600×600)	9	-	BBox, Segmentation, Label...
2022	LAION-5B	✓	Random Crawled	5.85B (-)	1	-	URL, Similarity, Language...
2022	DiffusionDB	✓	Synthetic Images	14M (-)	1	-	Size, Random Seed...
2022	COYO-700M	✓	Random Crawled	747M (-)	1	-	URL, Aesthetic Score...
2022	DeepFashion-MultiModal	✓	Full Body	44,096 (750×1101)	1	-	Densepose, Keypoints...
2023	ANNA	✓	News	29,625 (256×256)	1	-	-
2023	DreamBooth	✓	Objects & Pets	158 (-)	25	-	-

Inception Score (IS) 

Fréchet Inception Distance (FID) 

R-precision (RP) 

Semantic Object Accuracy (SOA) 


Positional Alignment (PA) 

## Generative Models


### ➤ GAN (2016~)

✓ Conditional GAN-based (Text-to-Face : 0/7)

[First work] Generative Adversarial Text to Image Synthesis

✓ StackGAN-based (Text-to-Face : 6/28)

[First work] StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks

✓ StlyeGAN-based (Text-to-Face : 8/10)

[First work] TediGAN: Text-Guided Diverse Image Generation and Manipulation

### ➤ Autogressive (2021~)

✓ Transformer-based

[First work] Zero-Shot Text-to-Image Generation

### ➤ Diffusion (2022~)

✓ Diffusion-based

[First work] High-Resolution Image Synthesis with Latent Diffusion Models

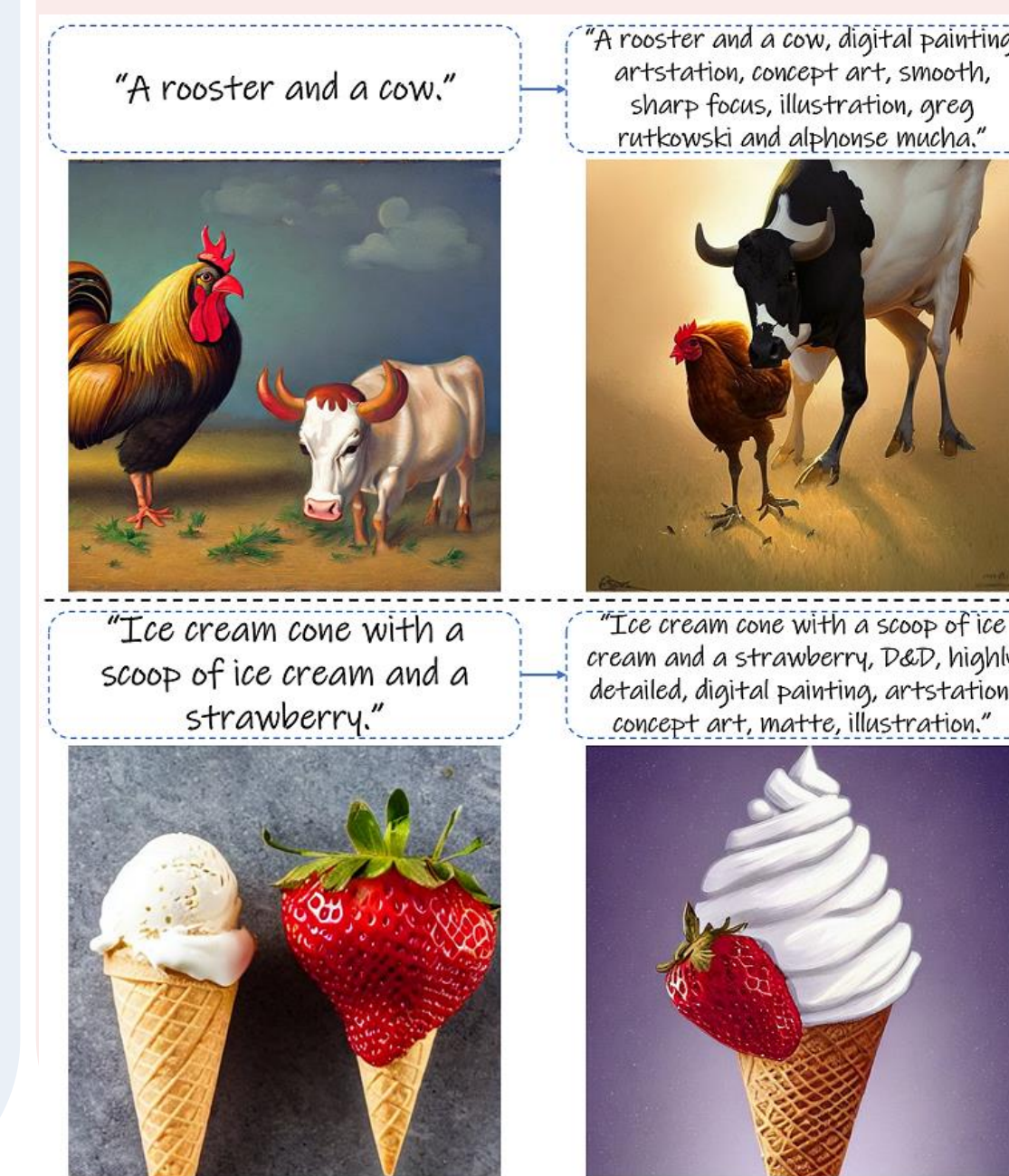
## Business Analysis

➤ **Computational Aesthetic** (Evaluation and Analysis)

➤ **Prompt Engineering**

➤ **Online Platforms**

➤ **Ethical Considerations**



Platform	Models	Price	Additional
Freehand	-	Free	Chinese prompts
Wombo	-	Free	Style
Craiyan	-	Free	-
Bing Image Creator	DALL-E 2	Free	-
SD Playground	Stable Diffusion	Free	-
Replicate	Stable Diffusion	Free	-
DeepAI	Stable Diffusion	Free	Style
HuggingFace	Stable Diffusion	Free	-
Yunjing	Stable Diffusion, ...	Free	Chinese prompts
Nightcafe	DALL-E 2, Stable Diffusion, ...	Free	Style
Lexica	-	Monthly 100 images	Search
starryai	-	Daily 5 free credits	Style, Image+Text
Dreamstudio	Stable Diffusion	200 free credits	Image+Text
Midjourney	-	20 times free	-
Firefly	-	Application is required	-
DALL-E 2	DALL-E 2	Monthly 15 free credits	-

## Generative Applications

### ➤ Text-to-Image

- Text-to-Face
- Text-to-Others

### ➤ Text-to-X

- Video/3D/Human Motion

### ➤ X-to-Image

- Text+Image/Layout
- Human Brain/Speed/Sound

### ➤ Multi Tasks



## Discussion

### ➤ Text-to-Face Task

- Existing text-to-face datasets suffer from a lack of large-scale image-text pairs
- Challenges
  - **Discriminability:** Recognizable as individual persons
  - **High Resolution & Photorealism:** Closely resemble authentic faces
  - **Fidelity:** Generated images are consistent with the input description.
  - **Controllability:** Selective manipulation with text prompts while preserving other irrelevant attributes.

### ➤ Text-to-X & X-to-image

- **Alignment:** Aligning modalities for accurate reflection of inputs.
- **Data scarcity:** Costly collection and annotation of large-scale multimodal datasets limit existing model performance.
- **Scalability:** Efficiently managing large-scale multimodal data in terms of memory and computational demands for multiple modalities.

### ➤ Universal Access & Commercial Use

### ➤ Versatile Models

- Reduce dependence on vast quantities of labeled data

Year	Method	Tasks										
		T2I	T2V	(T+X)2I	LYT2I	SKT2I	SEG2I	I2I	UIG	SR	IC	Other Tasks
2021	UFC-BERT	✓	-	Partial Image	-	-	-	✓	-	-	-	Multimodal Controls
2021	ERNIE-ViLg	✓	-	-	-	-	-	-	-	-	-	Generative VQA
2022	OFA	✓	-	-	-	-	-	-	-	-	-	VQA ...
2022	Frido	✓	-	-	-	-	-	-	-	-	-	SG2I
2022	LDMS	✓	-	-	✓	-	-	-	-	-	-	Inpainting
2022	NUWA	✓	-	Image/Video	-	-	-	✓	-	-	-	Video Prediction, ...
2022	MMVID	✓	-	Partial Image	-	-	-	-	-	-	-	Multimodal Controls
2022	PoE-GAN	✓	-	SEG/SKT/Image	✓	✓	-	-	-	-	-	(SEG+SKT)2I
2022	AugVAE-SL	✓	-	-	-	-	-	-	-	-	-	Image Reconstruction
2022	NUWA-Infinity	✓	✓	-	-	-	-	-	✓	-	-	Outpainting(HD), ...
2023	SDG	✓	-	Image	-	-	-	-	-	-	-	Style-guided, ...
2023	Muse	✓	-	Image	-	-	-	-	-	-	-	Inpainting, Outpainting
2023	MCM	✓	-	SEG/SKT	-	✓	-	-	-	-	-	(SEG+SKT)2I
2023	TextIR	✓	-	Image	-	-	-	-	-	-	-	Inpainting, Colorization
2023	GigaGAN	✓	-	Image	-	-	-	-	-	-	-	Style Mixing, ...
2023	UniDiffuser	✓	-	Image	-	-	-	-	✓	-	-	Joint Generation
2023	Visual ChatGPT	✓	-	Image	-	-	-	-	✓	-	-	Edge-to-Image, ...

<Acronym> Meaning: T2I:Text-to-Image; T2V:Text-to-Video; T+X:Text+X; LYT:Layout; SKT:Sketch; SEG:Segmentation; UIG:Unconditional Image Generation; SR:Super Resolution; IC:Image Captioning/Image-to-Text; VQA:Visual Question Answering; HD:High Definition; SG:Scene Graph.