
“A sophisticated AI lab filled with cutting-edge technology, where humanoid robots are being taught to paint like classical masters.” from Adobe Firefly (Beta)

# 画像生成AIの最先端

## 画像生成AIの進化と映像・3D生成への拡張

国立情報学研究所
コンテンツ科学研究系　池畑諭

# 私の提供する話題
# 画像生成AIの現状についての調査報告
＋映像生成＋3D生成

- 最先端のサービスが何かを知りたい
- 生成関連で何が研究近年盛んなのか知りたい

# 最近更新されたばかりの生成AIサービス

1 Midjourney 5.2（2023年6月） 技術詳細なし

2 Stable Diffusion XL（2023年6月） 技術詳細あり

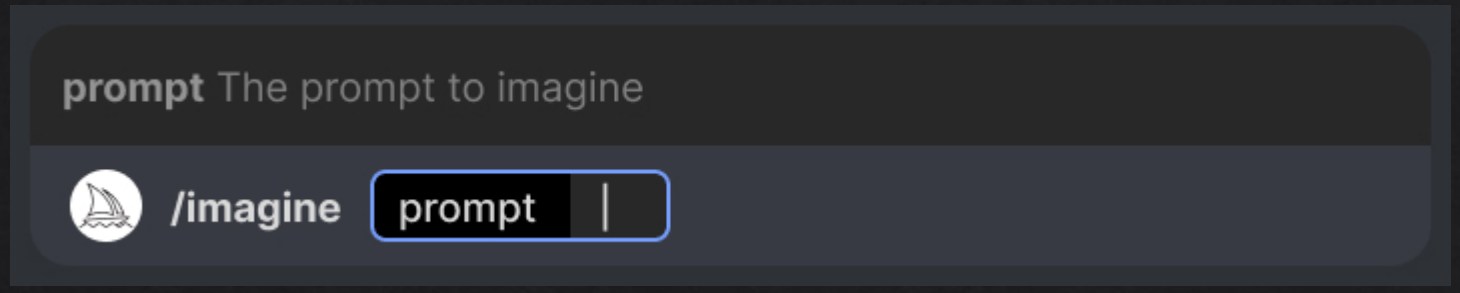
3 Adobe Firefly（2023年5月） 技術詳細なし

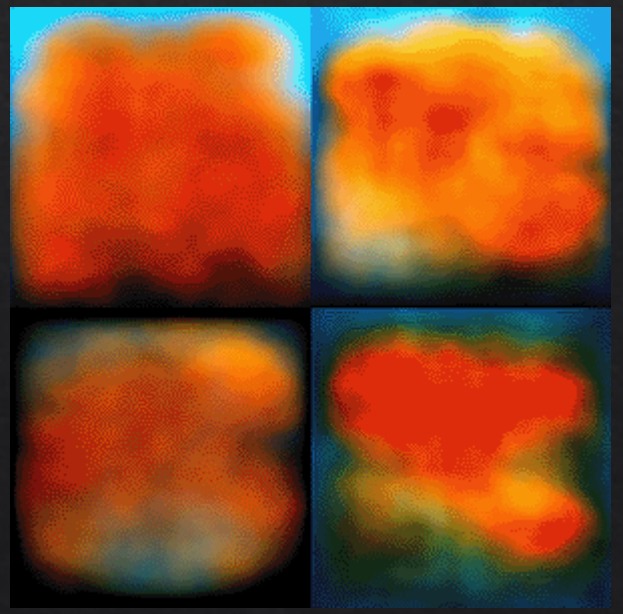
技術詳細なし

# 1 Midjourney

2022年にLeapMotion設立者のデビッドホルツ氏により設立（本社:米）

- $10, $30, $60, $120/月
  生成可能枚数や高速生成可能枚数などに違い
- 使い方はDiscordサーバー上でプロンプト入力
- 基本的には，生成した画像は誰でも見られる
  $60以上のプランに限りプライベート生成も可能



# 最新版 V5.2

より自然な照明効果と精細性の実現




“vivrant Califorlina flowers”

# 手の表現の進化

従来の生成モデルにおいて手の不自然さはたびたび指摘されてきた

V4(2022年11月)

V5.2(2023年6月)

https://www.reddit.com/r/midjourney/

https://www.youtube.com/watch?v=zB8Jr80HCnM&t=2s

# Zoom-Out

カメラ視点を遠ざけた画像を連続的に生成

# 2 Stable Diffusion

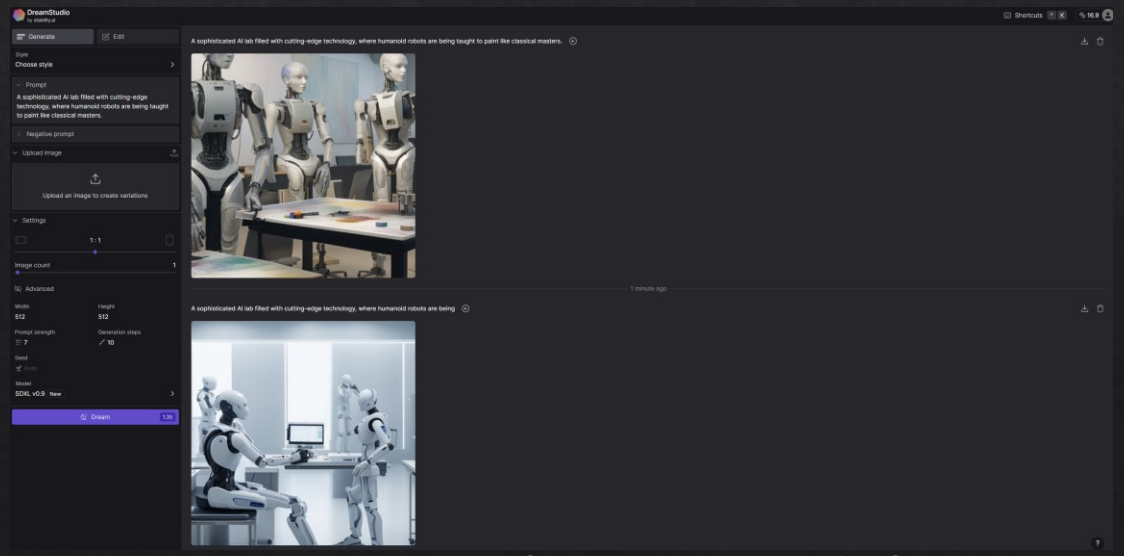
stability.ai

Latent Diffusion Model(CVPR2022)を元にstability.aiを設立(本社:英)

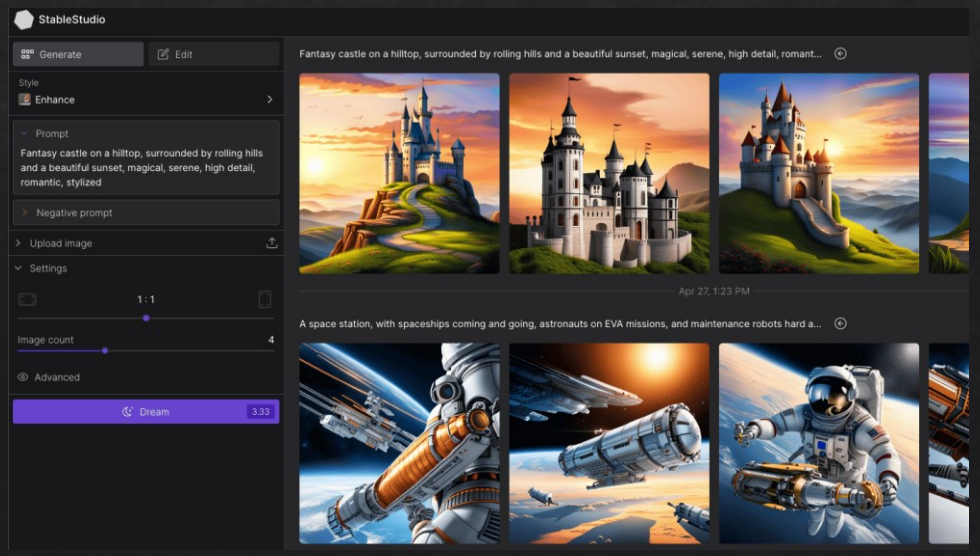
- オープンソース化されている (GitHub, Hugging Face等で無償利用可)
- 有料クラウドサービス(DreamStudio AI)も存在

  ポイント制 (1ポンド＝100Pt) 品質が並だと1枚1円程度, 最高設定では1枚で100円程度

  ✓ StableStudioとしてオープンソース化されている




DreamStudio (クラウドサービス)




StableStudio (オープンソース)

# Stable Diffusion XL（SDXL）

## 6月下順にSDXL 0.9を発表（7月中旬にオープンソース化）

“A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese”

SD 1.5（2022年10月）

SD 2.1（2022年12月）

最新版SDXL 0.9（2023年6月）

from StabilityAI Technical Report（2022/6）

stability.ai

# 2.1からSDXLへの主な変更点

- パラメータ数の増加 （8億→26億）
- より良いテキストモデル
  CLIP-ViT-L ＋ OpenCLIP ViT-bigG
- 学習スキームの改善
  256x256→512x512→1024x1024（異なるアスペクト比）
- 細部情報の高精度化モデルの導入

## *SDXL:* Improving Latent Diffusion Models for High-Resolution Image Synthesis


**Dustin Podell** **Zion English** **Kyle Lacey** **Andreas Blattmann** **Tim Dockhorn**

**Jonas Müller** **Joe Penna** **Robin Rombach**

Stability AI, Applied Research


*Code*: https://github.com/Stability-AI/generative-models *Model weights*: https://huggingface.co/stabilityai/



**Abstract**

We present *SDXL*, a latent diffusion model for text-to-image synthesis. Compared to previous versions of *Stable Diffusion*, *SDXL* leverages a three times larger UNet backbone: The increase of model parameters is mainly due to more attention blocks and a larger cross-attention context as *SDXL* uses a second text encoder. We design multiple novel conditioning schemes and train *SDXL* on multiple aspect ratios. We also introduce a *refinement model* which is used to improve the visual fidelity of samples generated by *SDXL* using a post-hoc *image-to-image* technique. We demonstrate that *SDXL* shows drastically improved performance compared the previous versions of *Stable Diffusion* and achieves results competitive with those of black-box state-of-the-art image generators. In the spirit of promoting open research and fostering transparency in large model training and evaluation, we provide access to code and model weights.

# 2.1からの改善点：より正確な文字表現

SD 1.5（2022年10月）

最新版SDXL 0.9（2023年6月）


“A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says "SDXL"!.”

# V.S. Midjourney 5.2

最新版V5.2（2023年6月）

最新版SDXL 0.9（2023年6月）


a green sign that says "Very Deep Learning" and is at the edge of the Grand Canyon

# Midjourney5.2 vs Stable Diffusion XL 0.9

## 照明の自然さ、”プロっぽさ” MJ5.2 > SDXL0.9

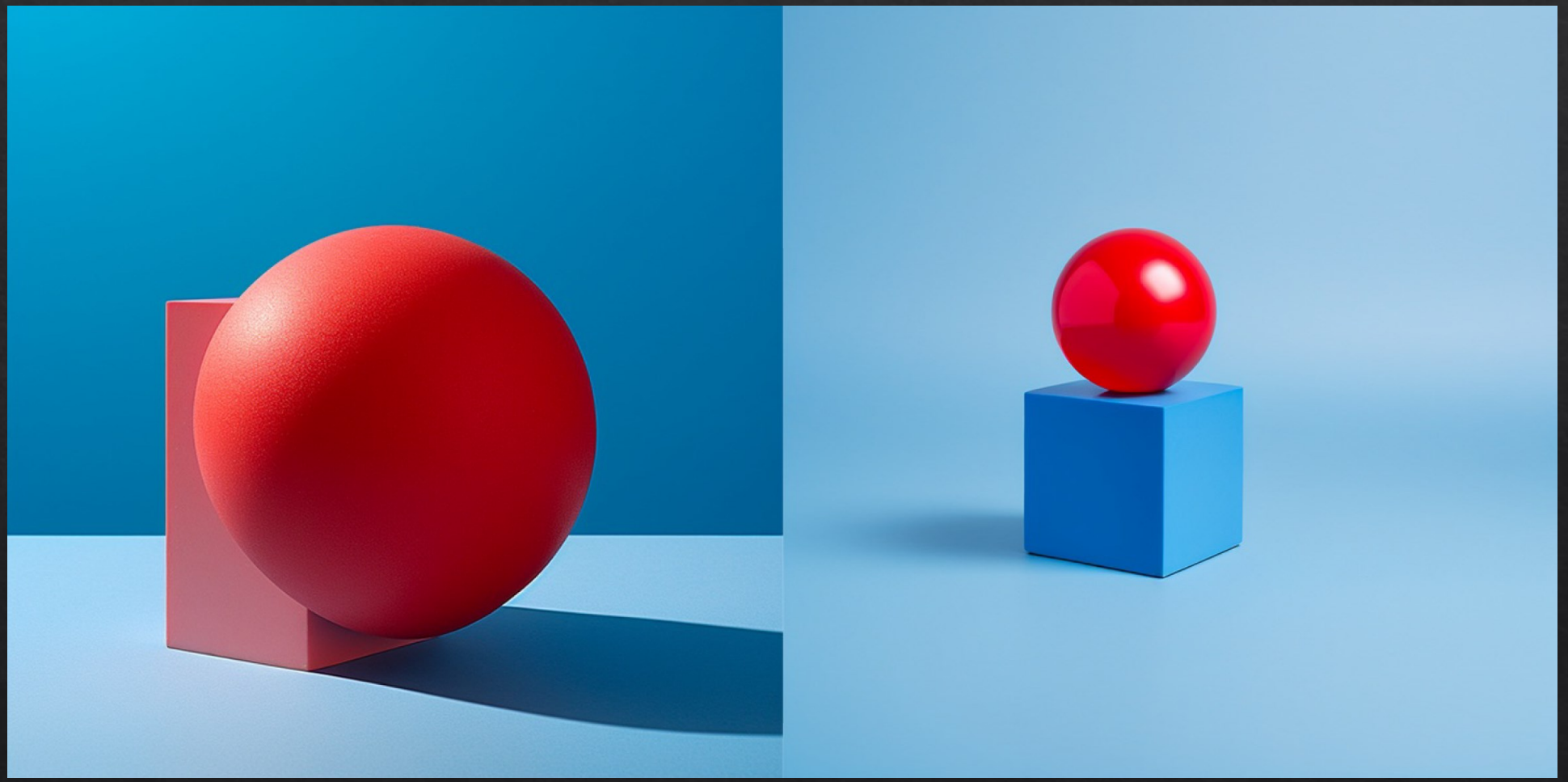

“A cinematic still of a close-up shot of a japanese ramen”



Image by Jim Clyde Monge

# Midjourney5.2 vs Stable Diffusion XL 0.9

## コンテクストの正しさ SDXL0.9 > MJ5.2



“photograph of a red ball on a blue cube”

# 3 Adobe Firefly

2023年3月クローズドβ、同5月にオープンβ


Firefly iguana having a great fashion sense

## 著者権クリーンなデータで学習

著作権切れ画像や自社AdobeStockの素材データのみ

## CAIによる生成物証明

Content Authenticity Initiative

## 不適切な生成の制御

暴力的なキーワードなどの規制

# 画像生成機能

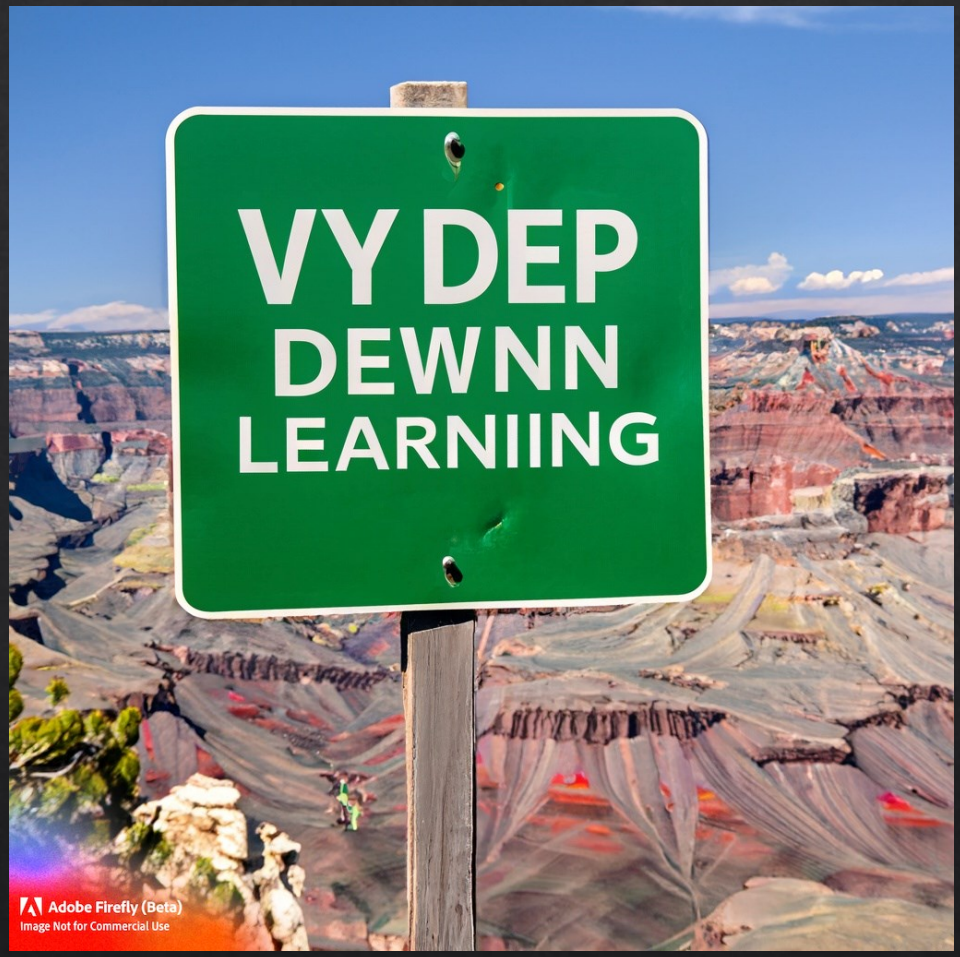
画像生成、インペインティング、テキストエフェクトの追加、色調補正はβ試用可能



“A close-up portrait of a black cat lounging leisurely”

+ タグ:pencil drawing

# 握手○　文字×

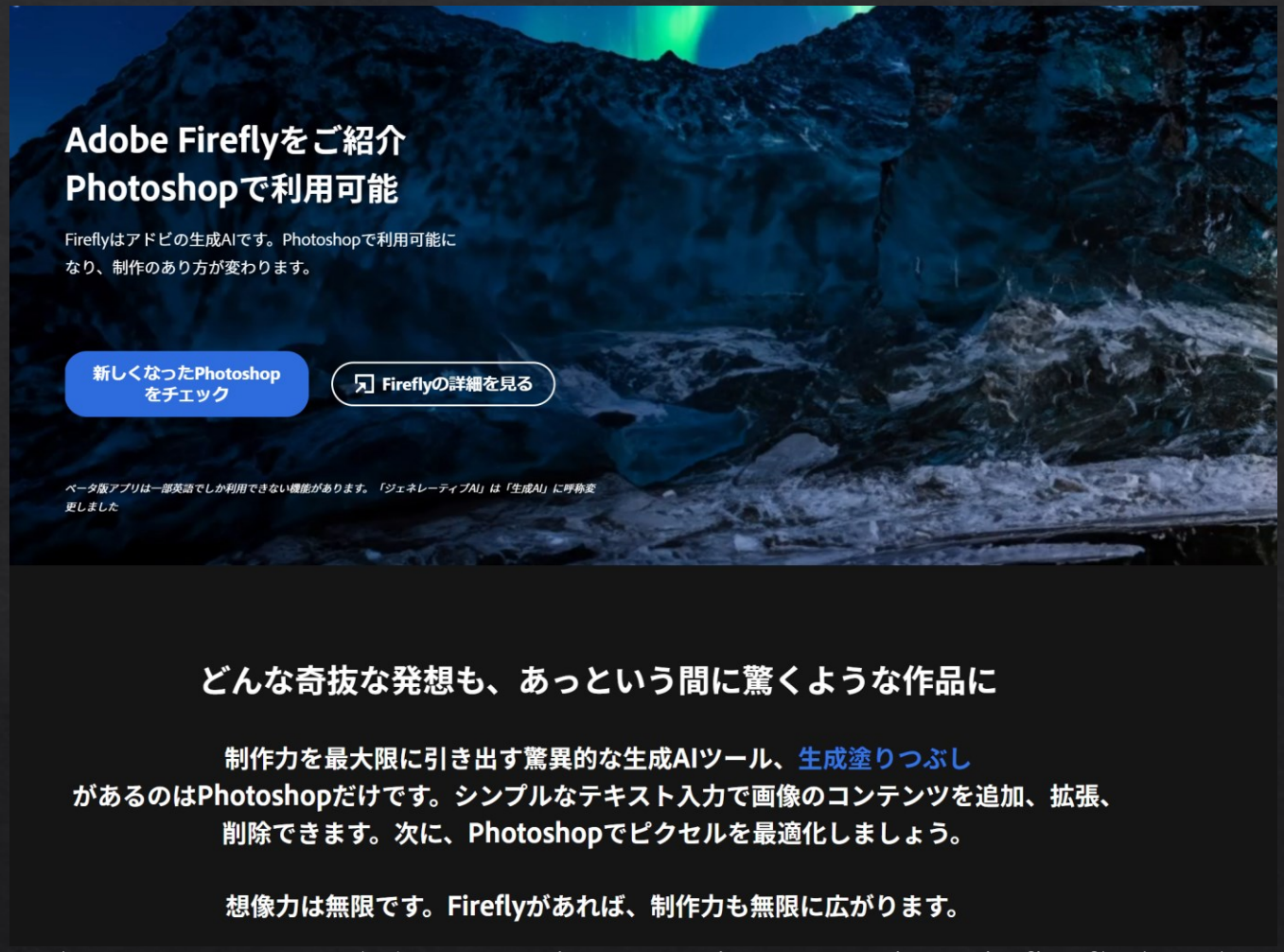




“A detailed photo of two individuals. They are in the middle of a firm handshake”

“a green sign that says "Very Deep Learning" and is at the edge of the Grand Canyon”

# Photoshop/Illustratorとの連携

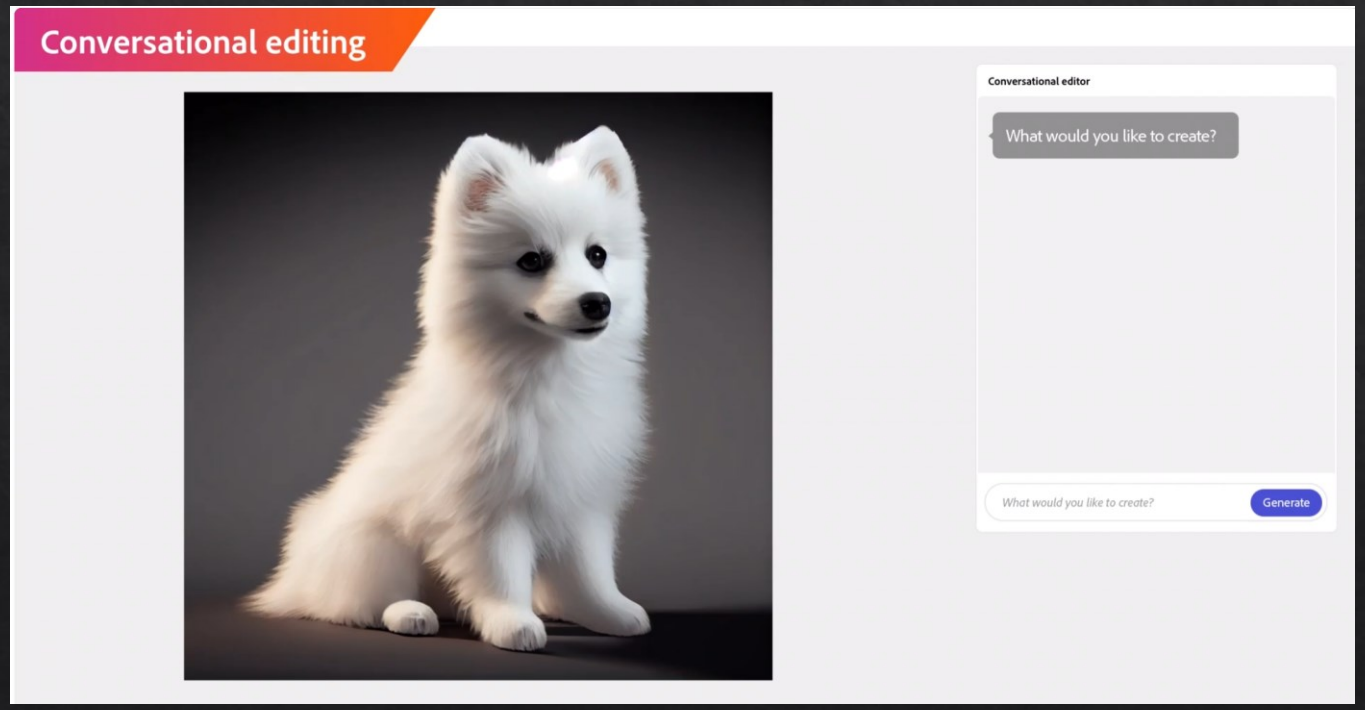
## 生成のみならず、編集やコンテンツクリエーションに特化




https://www.adobe.com/jp/sensei/generative-ai/firefly.html

# テキストに基づく画像編集

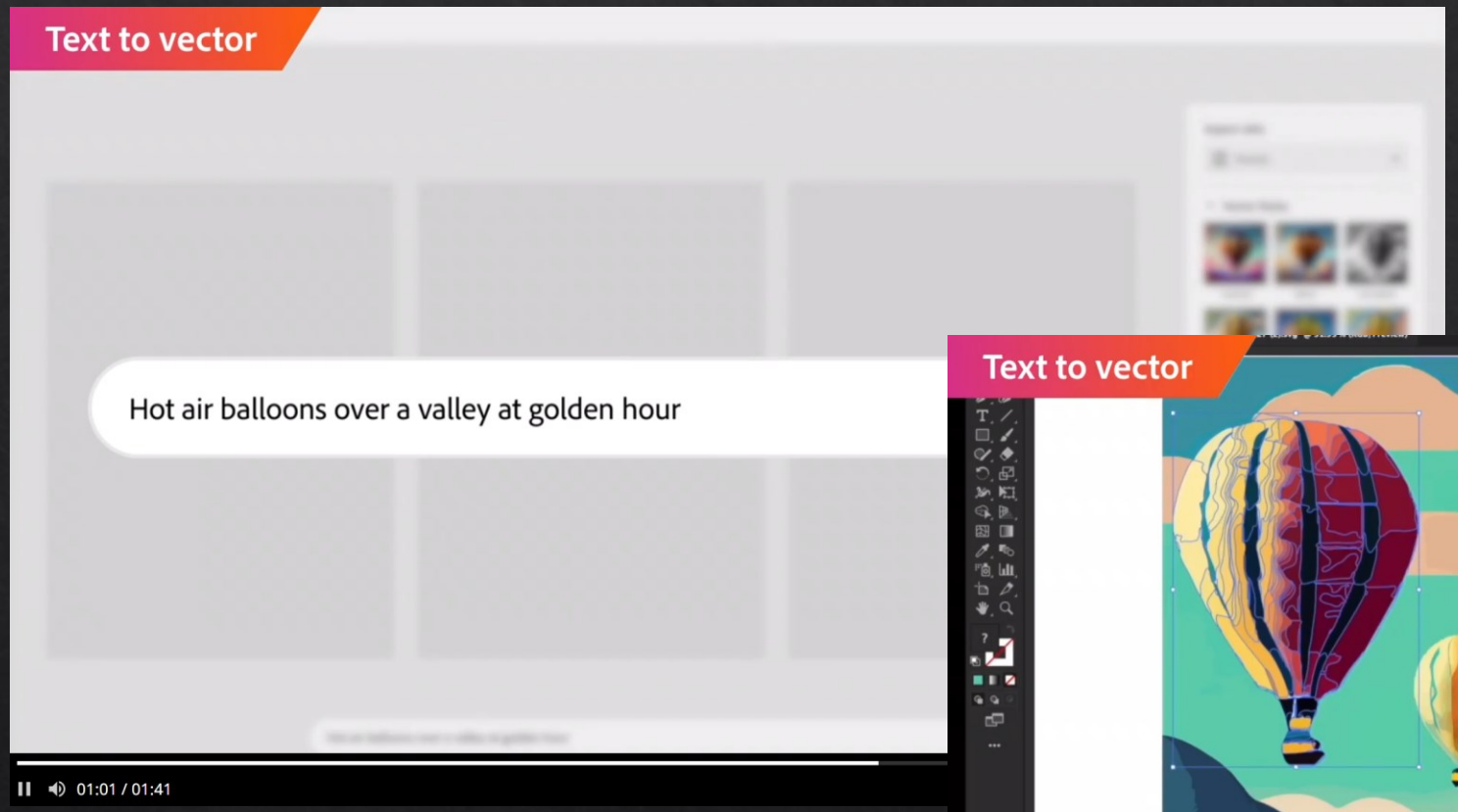
画像に映った「特定のもの」についての画像を生成する（Textual Inversion, Dreamboothなど研究が現在盛んにおこなわれている分野）




“Dress the dog in a Sanat suit
Put him in fron do a gingerbread house”

# ベクトル画像生成

ベクター画像の生成についての研究はCVPR2023等（VectorFusion）でも発表されている




# 画像生成AI　所感

- 明らかに研究段階からサービスに進んでいる（もちろん生成モデル自体の研究も盛んにおこなわれているが）
- 有償、無償を含めて生成物の品質はそこまで大差がない印象。独自性は機能面で持たせる必要がある
- 実際に使ってみて、プロンプトを考えるのが一番大変だと感じた。Chat AIがここにおいても有用性がある

# 映像生成AIの最先端

# 映像生成

映像生成の研究はGAN（Generative Adversarial Network）の時代から行われていた
（ただしテキスト条件付けではなく学習データでの条件付け）

Video GAN (Vondorick2016)


MoCoGAN（Tulyakov2017）

# Text-to-Videoモデル

| | | |
|---|---|---|
| Video Diffusion | Google 2022年3月<br>(arXiv) | 技術詳細あり |
| Make-A-Video | Meta 2022年9月<br>(ICLR2023) | 技術詳細あり |
| Imagen Video | Google 2022年10月<br>(arXiv) | 技術詳細あり |
| Gen-2 | Runway 2023年2月<br>6月7日から一般開放 | 技術詳細あり |
| Align your Latents | NVIDIA 2023年4月<br>(CVPR2023) | 技術詳細あり |
| (おまけ)Zeroscope | Hugginface 2023年6月 | 技術詳細なし |

動画は学習データが少ないのでtext-to-imageモデルをそのまま利用
それぞれの研究間で似たようなアプローチを共有している(Concurrentだらけ)

# Make-A-Video (META)

## Text-to-Video


A teddy bear painting a portrait

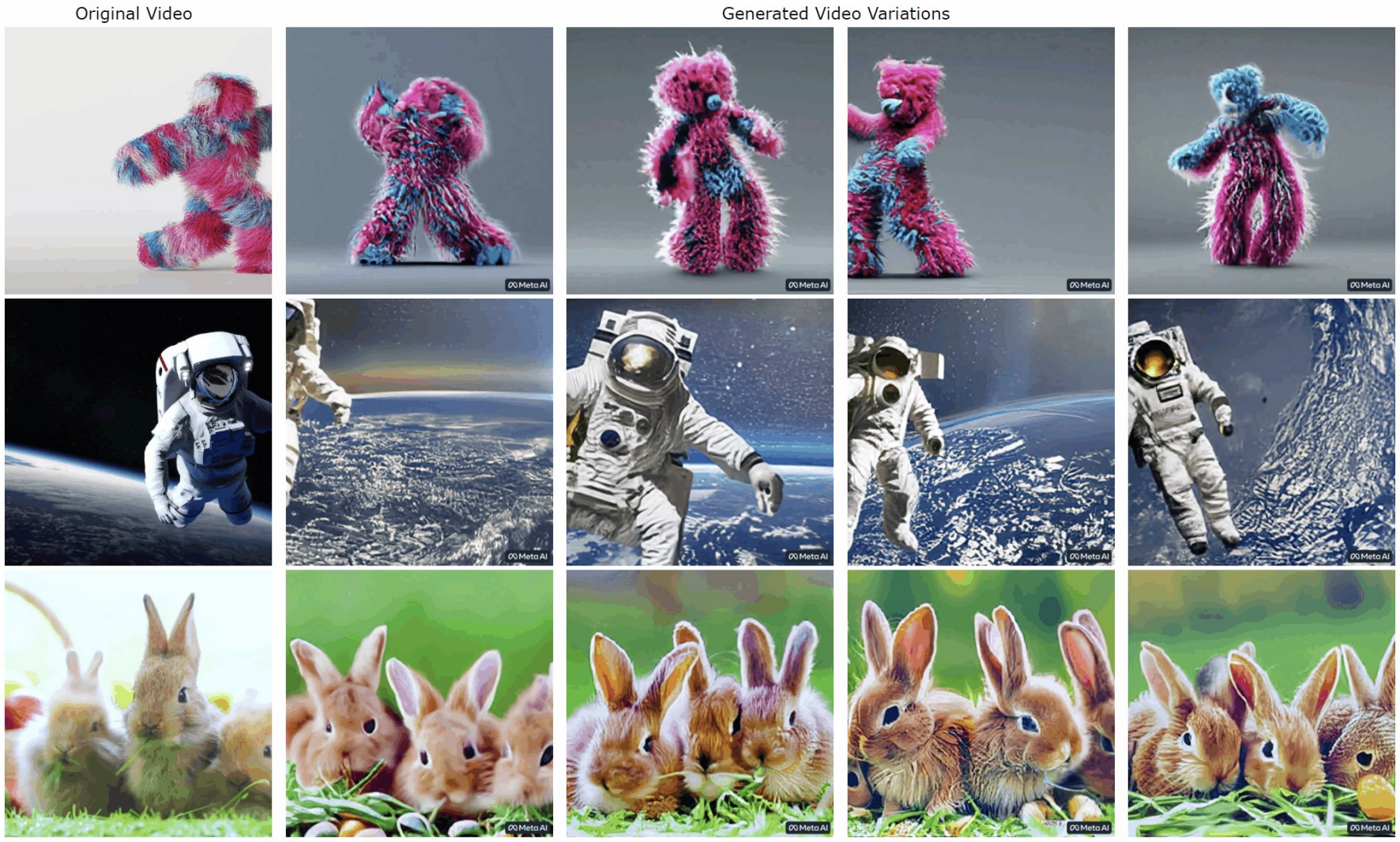

Clown fish swimming through the coral reef

# Make-A-Video（META）

静止画から動画

# Make-A-Video（META）

## ビデオのバリエーション生成



# Gen-2（Runway）

Text-to-Video（2023年6月一般利用可能に）


https://research.runwayml.com/gen2

# 映像生成AI　所感

- たった1年間での高画質化高精細化への進化が凄まじい
- しかし、映像生成AI特有の技術的なブレークスルーがあるわけでもないらしい（多くは画像生成の知見の直接的な流用）
- 根本的には単調な画像遷移を行っているだけで映像としてのストーリー性を持たせる事はまだ誰もできていない
- LLMを利用した映像への自動的なアノテーションを適切に行う事ができれば可能か？

# 3D生成AIの最先端

# テキストからの3次元生成


DreamFusion by Google Research
2022年9月

テキストを入力すると対応する3次元モデルが生成される

汎用的なゲームのアセット等に用いる応用が期待されている

# テキストからの3次元生成の難しさ

1. データセットが存在しない

テキストと3次元モデルのペアがそもそも存在しない。映像よりも希少

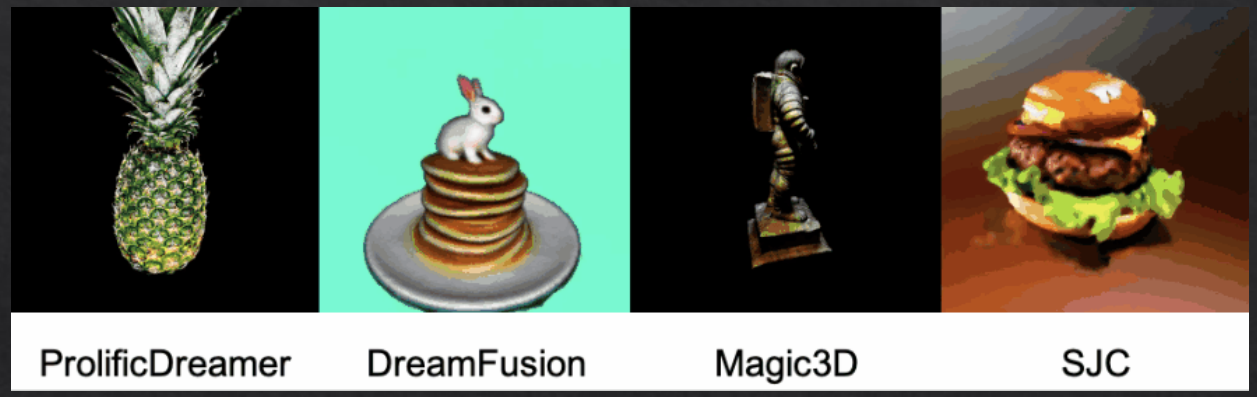
2. 3次元の生成モデルがほぼ存在しない

3次元の表現形式であるポリゴンやボリュームを直接扱うのはかなり困難

Text-to-imageモデル(特に拡散モデル)やNeRF(Neural Radiance Fields)との組み合わせで急速に発展している

# 3次元生成AIも群雄割拠

arXivでは毎月のように新しいtext-to-3Dのモデルの研究が発表されている

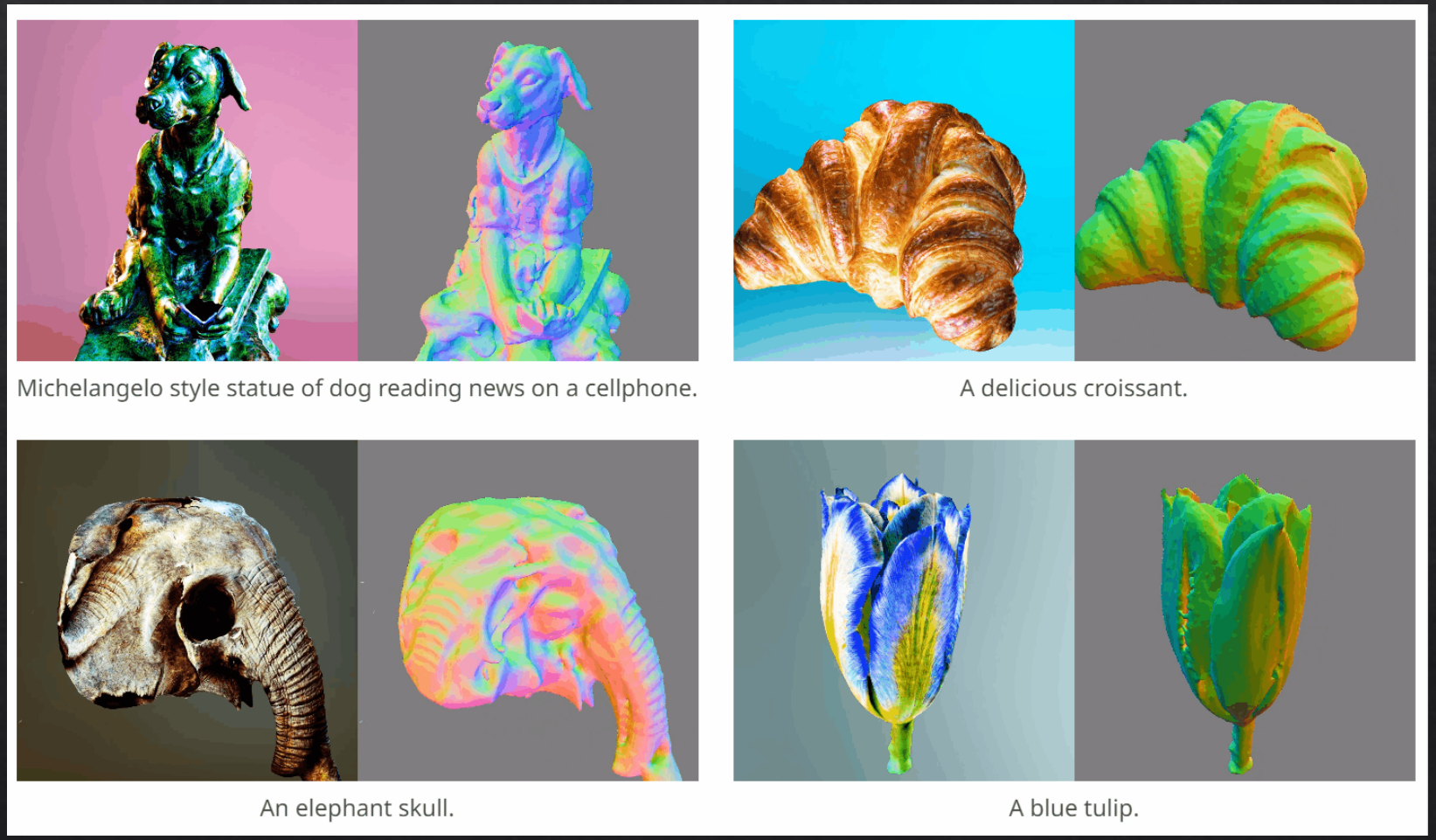



https://github.com/threestudio-project/threestudio

Make-A-Video4D（META, 2023年1月）

# 現状のText-to-3Dの最新モデル

オブジェクトレベルであれば写実的なテクスチャを生成可能（ただし、形状と材質の分離はまだほとんどできていない）




Prolific dreamer, Wang et al. 2023年５月. ArXiv

# Imagine 3D V1.2 (alpha)

An early experiment to prototype and create 3D with text
Access to generation is gradually expanding to everyone on the waitlist

search for a 3d model

search | join waitlist

erched blue jay bird, highly
etailed"
ratdreams_ai

"flying car"
@tahir

"Möbius strip"
@karanganesan

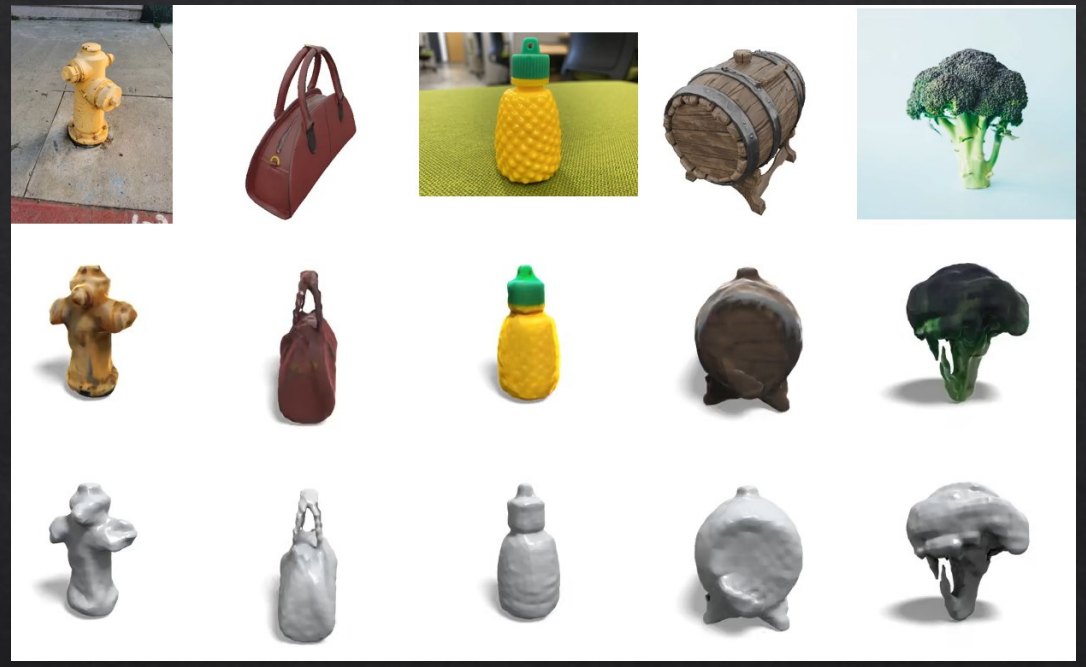
"Stylized beach hut"
@RicoUK

# Luma AI Imagin3D

# 1枚の画像から3次元モデル生成

3次元の復元と生成の複合的な課題で非常に挑戦的

Magic123, Qian et al. 2023年6月30日. arXiv



One-2-345, Liu et al. 2023年6月29日. arXiv

# 3次元生成AI　所感

- 映像同様に問題設定が定式化されてからの進歩が目覚ましい
- 拡散モデルとニューラルシーン表現というCV分野の近年における革新的な成果同士が結びついている
- 細かいディテールの制御や応用の有用性についてはまだサービスに至る段階ではないという印象

# 全体のまとめ

- 画像生成AIについては、サービス化が着々と進んでいる
- 映像生成AI、3次元生成AIについてはまだ研究段階という印象
- いわゆるビックテックは生成AIにどこも力を入れているが、アカデミアでもインパクトのある研究を行っている