# EvalCrafter：大規模ビデオ生成モデルのベンチマークと評価


**図1**：EvalCrafterは、テキストを動画に変換するモデルをベンチマークし評価するための包括的なフレームワークである。グレーで示された明確なプロンプトタイプと、黒丸で示された複数の評価項目を含む。

## 抄録

近年、視覚と言語の生成モデルは成長しすぎている。動画生成については、オープンソースで様々なモデルが公開されており、高画質な動画を生成するサービスも公開されている。しかし、これらの手法では、FVD [1] や IS [2] など、いくつかの学術的な指標を用いて性能を評価することが多い。我々は、大規模な条件付き生成モデルは、多くの場合、多面的な能力を持つ非常に大規模なデータセットで学習されるため、単純な指標から判断することは困難であると主張する。そこで我々は、生成されたビデオの性能を網羅的に評価するための新しいフレームワークとパイプラインを提案する。そのために、まず、大規模な言語モデルの助けを借りて実世界のプロンプトリストを分析することにより、テキスト-ビデオ生成のための新しいプロンプトリストを実施する。次に、私たちが慎重に設計したベンチマークで、映像の質、コンテンツの質、動きの質、テキストとキャプションの整合性の観点から、最先端のビデオ生成モデルを約18の客観的メトリクスで評価する。モデルの最終的なリーダーボードを得るために、我々はまた、客観的な指標をユーザーの意見に合わせるために、一連の係数をフィットさせた。提案されたオピニオンアライメント手法に基づき、最終的なスコアは、単に評価指標を平均化するよりも高い相関を示し、提案された評価手法の有効性を示している。

## 1 はじめに

大規模なジェネレーティブ・モデルの魅力が世界を席巻している。例えば、よく知られている ChatGPT と GPT4 [3] は、コーディング、数学の問題を解くこと、さらには視覚的な理解など、いくつかの側面において人間レベルの能力を示しており、これは会話形式であらゆる知識を用いて人間と対話するために使用することができる。ビジュアルコンテンツ作成のための生成モデルとしては、Stable Diffusion [4] と SDXL [5] が非常に重要な役割を果たしている。

テキストから画像への変換にとどまらず、動画生成のための拡散モデルの飼いならしも急速に進んでいる。初期の作品（ImagenViedo [6]、Make-A-Video [7]）は、カスケードモデルを直接ビデオ生成に利用している。Stable Diffusion における画像生成の事前分布を利用した LVDM [20] と MagicVideo [69] が、効率的に動画を生成するための時間的レイヤーを学習するために提案されている。学術論文とは別に、いくつかの商用サービスもテキストや画像からビデオを生成することができる。例えば、Gen2 [8] や PikaLabs [9] などである。これらのサービスの技術的な詳細は分からないが、他の方法との評価や比較はされていない。しかし、現在の大規模な T2V（text-to-video）モデルはすべて、FVD [1:1] のような以前の GAN ベースのメトリクスを評価に使用しているだけであり、テキストプロンプトと生成されたビデオ間のペア以外の、生成されたビデオと実際のビデオ間の分布マッチングにのみ関係している。これとは異なり、我々は、優れた評価方法は、異なる側面、例えば、動きの質や時間的一貫性などのメトリクスを考慮すべきだと主張している。また、大規模な言語モデルと同様に、公開されていないモデルもあり、生成されたビデオにしかアクセスできないため、評価の難しさがさらに増している。LLM [3:1]、MLLM [10]、text-to-image [11] など、大規模な生成

モデルの分野では評価が急速に進んでいるが、これらの手法を映像生成に直接利用することはまだ難しい。ここでの主な問題は、テキストから画像への変換や対話の評価とは異なり、動きと一貫性がビデオ生成にとって非常に重要であることである。

我々は、ビデオのための大規模なマルチモダリティ生成モデルを評価する最初のステップを行う。具体的には、まず、様々な日常的なオブジェクト、属性、動作を含む包括的なプロンプトリストを作成する。よく知られた概念のバランスのとれた分布を実現するために、我々は実世界の知識のよく定義されたメタタイプから出発し、ChatGPT [3:2] などの大規模な言語モデルの知識を利用して、我々のメタプロンプトを広範囲に拡張する。モデルによって生成されたプロンプトの他に、実世界のユーザーからのプロンプトとテキストから画像へのプロンプトも選択する。その後、さらなる評価用途のために、プロンプトからメタデータ（色、サイズなど）も取得する。第二に、これらの大規模なT2Vモデルの性能を、映像の視覚的品質、テキストと映像の整合性、動きの品質と時間的整合性など、さまざまな側面から評価する。各アスペクトについて、1つ以上の客観的な指標を評価指標として使用する。これらの指標はモデルの能力の1つを反映しているに過ぎないので、モデルの資質を判断するために、多面的なユーザー調査も行っている。

全体として、我々はこの論文の貢献を次のように要約する。

- 我々は、まず大規模な T2V モデルを評価し、T2V 評価のための詳細なアノテーションを含む包括的なプロンプトリストを構築する。
- 我々は、ビデオ生成の評価のために、ビデオの視覚的品質、ビデオの動きの品質、およびテキストとビデオの整列の側面を考慮する。各アスペクトについて、人間の意見を整列させ、また人間の整列によって提案メトリックの有効性を検証する。
- T2V 生成モデルのさらなるトレーニングに役立つかもしれない、いくつかの結論と発見についても議論した。

# 2 関連研究

## 2.1 Text-to-Video 生成と評価

T2V の生成は、与えられたテキストプロンプトからビデオを生成することを目的としている。初期の研究では、変分オートエンコーダ（VAE [12]）または生成的敵対ネットワーク（GAN [13]）によってビデオを生成している。しかし、生成される動画の質は、しばしば低品質であったり、顔 [14] や風景 [15] [16] など、特定の領域でしか機能しなかったりする。拡散モデル [17]、ビデオ拡散モデル [18]、大規模なテキスト-画像事前学習 [19] の急速な発展に伴い、現在の方法は、生成前に、より強力なテキスト-画像事前学習モデルを利用する。例えば、Make-A-Video [7:1] や Imagen-Video [6:1] は、カスケードされたビデオ拡散モデルを学習し、いくつかのステップでビデオを生成する。LVDM [20]、Align Your latent [21]、MagicVideo [22]は、時間的な注目やトランスフォーマー層を追加することで、潜在的なテキスト-画像モデルをビデオ領域に拡張している。AnimateDiff [23] は、パーソナライズされたテキスト-画像モデルを利用することで、良好な視覚的品質を示す。同様の手法は SHOW-1 [24]や LAVIE [25]でも提案されている。T2V 生成は、コマース企業や非コマース企業の熱意も高める。Gen1 [8:1] や Gen2 [8:2] などのオンラインモデルサービスでは、完全な T2V 生成や条件付きビデオ生成において、高品質な生成ビデオの能力を示している。Discord ベースのサーバーでは、Pika-Lab [9:1]、Morph Studio [26]、FullJourney [27]、Floor33 Pictures [20:1] も非常に競争力のある結果を示している。さらに、ZeroScope [28]、ModelScope [29] など、オープンソースのテキスト（または画像）-動画モデルもある。

しかし、これらの方法には、それぞれの方法の利点を評価するための公正で詳細なベンチマークがまだ欠けている。例えば、FVD [1:2]（LVDM [20:2]、MagicVideo [22:1]、Align Your Latent [21:1]）、IS [2:1]（Align Your Latent [21:2]）、CLIP 類似性 [19:1]（Gen1 [8:3]、Imagen Video [6:2]、Make-A-Video [7:2]）、またはパフォーマンスレベルを示すユーザースタディを使用してパフォーマンスを評価するだけである。これらの指標は、これまでの領域内テキスト画像生成手法に対してのみ有効で、T2V 生成にも重要な、入力テキストのアライメント、動きの質、時間的整合性を無視している可能性がある。

## 2.2 大規模生成モデルの評価

大規模な生成モデル [3:3] [5:1] [4:1] [30] [31] を評価することは、自然言語処理と視覚タスクの両方にとって大きな課題である。大規模な言語モデルのために、現在のメソッドは、異なる能力、質問タイプ、およびユーザーのプラットフォーム [32] [33] [34] [35] [36] の観点からいくつかのメトリックを設計する。LLM 評価とマルチモデルLLM評価の詳細については、最近の調査 [37] [38] を参照されたい。同様に、マルチモーダル生成モデルの評価も研究者の注目を集めている [39] [40]。例えば、Seed-Bench [10:1] は、マルチモーダルな大規模言語モデル評価のための VQA を生成する。

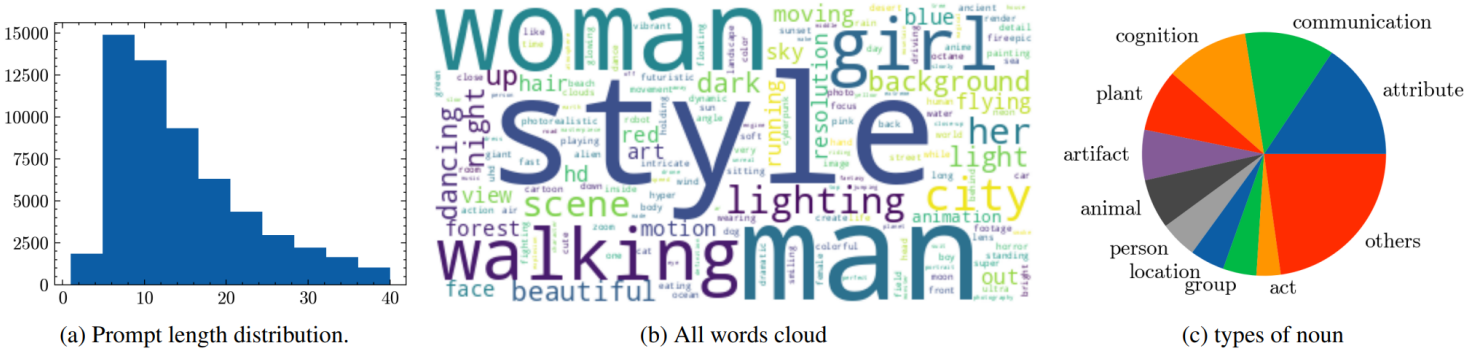
ビジュアル生成タスクのモデルについては、Imagen [41] はユーザー調査によってのみモデルを評価している。DALL-Eval [42] は、ユーザとオブジェクト検出アルゴリズム [43] の両方を介して、テキスト画像モデルの視覚的推論スキルと社会的基盤を評価する。HRS-Bench [44] は、ChatGPT [3:4] を用いてプロンプトを生成し、テキスト画像モデルの13のスキルを評価するために17のメトリックを利用することにより、全体

的で信頼性の高いベンチマークを提案する。TIFA [11:1] は、視覚的質問応答(VQA)を利用したベンチマークを提案している。しかし、これらの方法は、テキストから画像への評価や言語モデルの評価には依然として有効である。T2V 評価では、動きの質と時間的一貫性を考慮する。

# 3 ベンチマーク構成

我々のベンチマークは、T2V の様々なモデルの能力を公平に評価するために、信頼できるプロンプトリストを作成することを目的としている。この目標を達成するために、我々はまず、大規模な実世界のユーザーから T2V プロンプトを収集し、分析する。その後、生成されたプロンプトの多様性を高めるための自動パイプラインを提案し、事前に訓練されたコンピュータビジョンモデルによって識別・評価できるようにする。ビデオ生成には時間がかかるため、初期バージョンとして500のプロンプトを収集し、入念なアノテーションを施して評価する。以下、各ステップの詳細を説明する。

## 3.1 どのようなプロンプトを作成すべきか




(a) Prompt length distribution. (b) All words cloud (c) types of noun

**図2**：PikaLab Server [9:2] の実世界プロンプトの分析。

この問いに答えるために、我々は、FullJourney [27:1] や PikaLab [9:3] を含む、実際の T2V 生成の discord ユーザーからプロンプトを収集する。合計で60万以上のプロンプトと対応するビデオを取得し、繰り返されるプロンプトや無意味なプロンプトを削除することによって20万にフィルタリングする。図2(a) に示すように、プロンプトの 90% は $[3, 40]$ の範囲の単語を含んでいる。また、図2(b) では、人物、スタイル、人間の動き、シーンが支配的で、ビデオ、カメラ、高い、品質などの不明瞭な単語を削除して、最も重要な単語をプロットしている。上記の分析にもかかわらず、プロンプトリストのメタクラスを決定するために単語クラスもカウントする。図2(c) に示すように、WordNet [45] を用いてメタクラスを同定すると、コミュニケーション、属性、認知の各単語を除いて、不自然さ（人造物）、人間、動物、場所（景観）が重要な役割を果たす。また、図2(b) の最も重要な単語スタイルをメタクラスに追加する。全体として、我々は T2V 生成を、人間、動物、物体、風景を含む、およそ4つのメタ被写体クラスに分けた。また、それぞれのタイプについて、モーションやスタイル、現在のメタクラスと他のメタクラスとの関係も考慮し、映像を構成する。さらに、主オブジェクトに関連し、ビデオにとって重要なモーションを含める。最後に、カメラの動きとテンプレートによるスタイルを検討する。

## 3.2 一般的な認識可能プロンプト生成

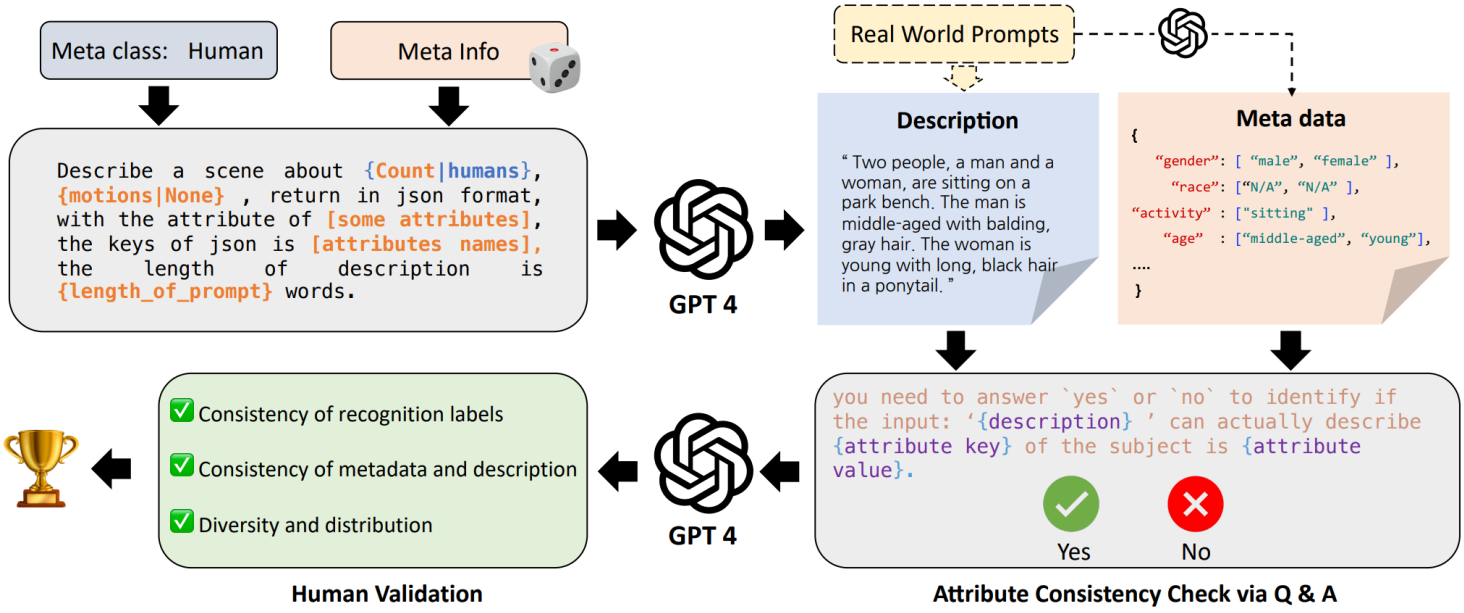



**図3**：我々は、コンピュータビジョンモデルとユーザーによるテキストからビデオへの評価のために、詳細なプロンプトを備えた信頼できるベンチマークを生成することを目指している。そのパイプラインを上に示す。

### 自動プロンプト生成

プロンプトリストのメタクラスを決定した後、大規模言語モデル（LLM）と人間の力によって認識可能なプロンプトを生成する。図3に示すように、メタクラスの種類ごとに、GPT-4 [3:5] に、このメタクラスに関するシーンを、シーンの属性とともにランダムにサンプリングしたメタ情報で記述させることで、すでにラベルがわかっているようにする。例えば人間の場合、GPT-4 に人間の属性、年齢、性別、服装、人間の活動などを教えてもらうことができ、これらは JSON ファイルとしてグランドトゥルースコンピュータビジョンモデルとして保存される。しかし、GPT-4 はこのタスクに対して完全ではなく、生成された属性は生成された記述とあまり一致していないこともわかった。そこで、生成された説明文とメタデータの類似性を識別するために GPT-4 も使用し、ベンチマーク構築にセルフチェックを組み込む。最後に、各プロンプトが正しく、T2V 生成に意味のあるものであることを確認するために、私たち自身でプロンプトをフィルタリングする。

### 現実世界からのプロンプト

我々はすでに実世界のユーザーから非常に大規模なプロンプトを収集しており、DALL-Eval [42:1]やDraw-Bench [41:1]など、利用可能なテキストから画像への評価プロンプトもあるので、これらのプロンプトもベンチマークリストに統合する。これを実現するために、まずGPT-4を使ってフィルタリングし、メタデータを生成する。そして、図3 に示すように、対応するメタ情報とともに適切なプロンプトを選択し、メタ情報の整合性をチェックする。

### 3.3 ベンチマーク分析

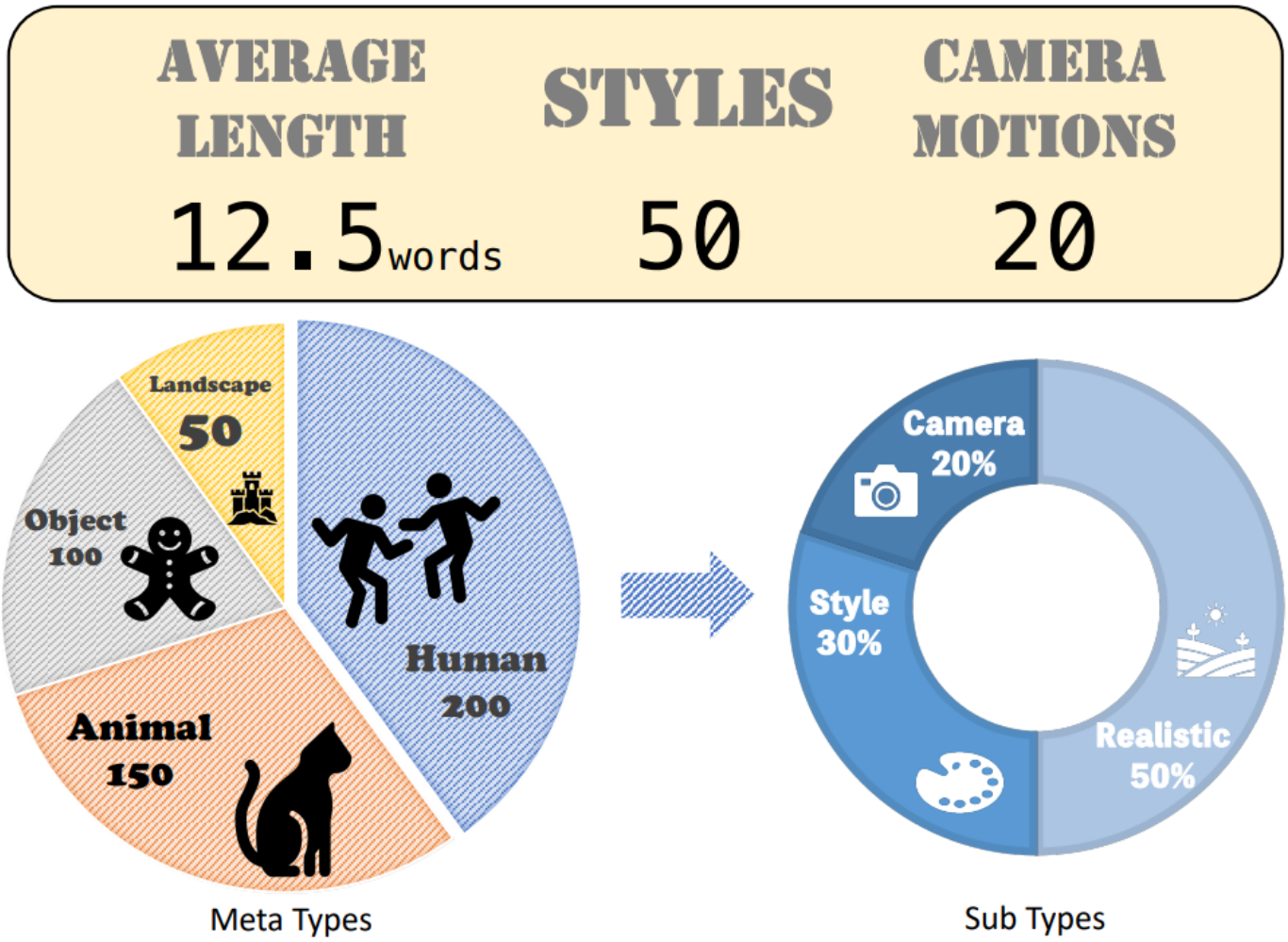



**図4**：**提案されたベンチマークの分析**。各メタタイプには3つのサブタイプがあり、生成される動画の多様性を高めている。

全体として、人間、動物、物体、風景のメタクラスで500以上のプロンプトが得られる。各クラスには、自然なシーン、定型化されたプロンプト、明示的なカメラモーションコントロールによる結果が含まれる。図4 にベンチマークの概要を示す。ベンチマーク全体では、500以上のプロンプトが慎重に分類されている。プロンプトの多様性を高めるために、我々のベンチマークには図4 に示すように3つの異なるサブタイプがあり、合計50のスタイルと20のカメラモーションプロンプトがある。全ベンチマーク中50%のプロンプトにランダムに追加する。私たちのベンチマークでは、平均12.5語の事前プロンプトが含まれており、これは図2 に見られるように、実際のプロンプトと同様である。

## 4 評価指標

| Method | Ver. | Abilities† | Resolution | FPS | Open Source | Length | Speed | Motion | Camera |
|---|---|---|---|---|---|---|---|---|---|
| ModelScope | 23.30 | T2V | 256 × 256 | 8 | ✓ | 4*s* | 0.5 min | – | – |
| VideoCrafter | 23.04 | T2V | 256 × 256 | 8 | ✓ | 2*s* | 0.5 min | – | – |
| | | | | | | | | | |
| ZeroScope | 23.06 | T2V&V2V | 1024 × 576 | 8 | ✓ | 4*s* | 3 min | – | – |
| ModelScope-XL | 23.08 | I2V&V2V | 1280 × 720 | 8 | ✓ | 4*s* | 8 min+ | – | – |
| Floor33 Pictures | 23.08 | T2V | 1280 × 720 | 8 | – | 2*s* | 4 min | – | – |
| PikaLab | 23.09 | I2V or T2V | 1088 × 640 | 24 | – | 3*s* | 1 min | ✓ | ✓ |
| Gen2 | 23.09 | I2V or T2V | 896 × 512 | 24 | – | 4*s* | 1 min | ✓ | ✓ |

**表1**：**利用可能な拡散ベースのテキストからビデオへのモデルの違い**。† テキストからビデオへの変換（T2V）の方法を主に評価する。関連する画像-動画生成モデル（I2V）、すなわちModelScope-XLでは、まずStable Diffusion v2.1によって画像を生成し、生成されたコンテンツに対して画像-動画を実行する。

これまでの FID [46] に基づく評価指標とは異なり、生成された映像の視覚的品質、テキストと映像の整合性、コンテンツの正しさ、動きの品質、時間的整合性など、さまざまな側面から T2V モデルを評価する。以下に詳細な指標を示す。

## 4.1 総合的なビデオ品質評価

私たちはまず、生成されたビデオのビジュアルクオリティを考慮する。分布に基づく手法、例えば FVD [1:3] は、評価のために依然としてグランドトゥルース映像を必要とするため、これらの種類のメトリクスは、一般的な T2V 生成のケースには適していないと主張する。

### ビデオ品質評価（VQA$_A$, VQA$_T$）

生成されたビデオの品質を美的および技術的に評価するために、最先端のビデオ品質評価手法である Dover [47] を利用する。技術的評価では、ノイズや不自然さなどの一般的な歪みの観点から生成されたビデオの品質を測定する。Dover [47:1] は、自己収集されたより大規模なデータセットで学習され、ラベルはアライメントのために実際のユーザーによってランク付けされる。美的スコアはVQA$_A$、技術的スコアはVQA$_T$と呼ぶ。

### インセプションスコア（IS）

また、T2V 生成の先行論文のメトリクスに倣い、映像のインセプションスコア [2:2] も映像品質評価指標の1つとして用いる。インセプションスコアは GAN [13:1] の性能を評価するために提案されたもので、ImageNet [48] データセット上で事前に訓練された特徴抽出手法として、事前に訓練されたインセプションネットワーク [49] を利用する。インセプションスコアは生成された動画の多様性を反映し、スコアが大きいほど生成されたコンテンツの多様性が高いことを意味する。

## 4.2 テキストとビデオの位置合わせ

もう一つの一般的な評価方向は、入力テキストと生成されたビデオの位置合わせである。私たちは、グローバルなテキストプロンプトとビデオの両方を考慮するだけでなく、さまざまな側面からコンテンツの正しさも考慮します。以下に各スコアの詳細を記す。

### テキストとビデオの整合性（CLIP-Score）

CLIP-Score は、入力されたテキストプロンプトと生成されたビデオとの間の不一致を定量化するために広く使用され、簡便であることから、評価指標の1つとして取り入れた。事前に学習された ViT-B/32 CLIP モデル [19:2] を特徴抽出器として利用し、フレーム単位の画像埋め込みとテキスト埋め込みを求め、それらの余弦類似度を計算する。$i$ 番目のビデオ $x_t^i$ の $t$ 番目のフレームと対応するプロンプト $p^i$ の余弦類似度は、$emb(\cdot)$ を CLIP エンベッディングとして、$\mathcal{C}\left(emb\left(x_t^i\right), emb\left(p^i\right)\right)$ と表せる。総合 CLIP-Score（$S_{CS}$）は、すべてのフレームとビデオにわたる個々のスコアの平均によって導き出され、

$$S_{CS} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N}\sum_{t=1}^{N}\mathcal{C}\left(emb\left(x_t^i\right), emb\left(p^i\right)\right)\right) \tag{1}$$

のように計算される。ここで、$M$ はテストビデオの総数、$N$ は各ビデオのフレームの総数である。

### テキストとビデオの整合性（SD-Score）

現在のビデオ拡散モデルのほとんどは、より大規模なデータセットを用いて、ベースとなる安定した拡散を微調整したものである。また、Stable Diffusion の新しいパラメータをチューニングすることは、概念的な忘却を引き起こすため、生成された品質をフレーム単位の Stable Diffusion [4:2] と比較することで、新しい指標を提案する。具体的には、SDXL [5:2] を用いて、プロンプトごとに $N_1$ 枚の画像 $\{dk\}_{k=1}^{N_1}$ を生成し、生成された画像とビデオフレームの両方から視覚的埋め込みを抽出する。ここで、我々は $N_1 = 5$ とした。生成された動画と SDXL 画像の埋め込み類似度を計算する。これは、テキストから画像への拡散モデルを動画モデルに微調整する際に、概念忘れの問題を解消するのに役立つ。最終的なSDスコアは以下のようになる。

$$S_{SD} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N}\sum_{t=1}^{N}\left(\frac{1}{N_1}\sum_{k=1}^{N_1}\mathcal{C}\left(emb\left(x_t^i\right), emb\left(d_k^i\right)\right)\right)\right) \tag{2}$$

### テキストとビデオの整合性（BLIP-BLEW）

また、生成されたビデオのテキスト説明と入力テキストプロンプトとの間の評価も考慮する。この目的のために、キャプション生成に BLIP2 [50] を利用する。テキストから画像への評価方法 [44:1] と同様に、フレーム間で生成されたプロンプトとソースプロンプトのテキストアライメントに BLEU [51] を使用する。

$$S_{BB} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N_2}\sum_{k=1}^{N_2}\mathcal{B}\left(p^i, l_k^i\right)\right) \tag{3}$$

ここで、$\mathcal{B}(\cdot,\cdot)$ は BLEU 類似度スコアリング関数であり、$\{l_k^i\}_{k=1}^{N_2}$ は $i$ 番目のビデオに対して BLIP が生成したキャプションであり、$N_2$ は実験的に $5$ に設定されている。

### オブジェクトと属性の整合性（Detection-Score, Count-Score および Color-Score）

一般的なオブジェクトに対しては、最先端のセグメンテーションとトラッキング手法である SAM-Track [52] を採用し、我々が関心を持つビデオコンテンツの正しさを分析する。強力なセグメンテーションモデル [53] を活用することで、オブジェクトとその属性を簡単に得ることができる。我々のパイプラインでは、COCO クラス [54] を持つプロンプトの検出に焦点を当てる。COCO クラスは、オブジェクト検出とセグメンテーションタスクに広く使用されているデータセットである。T2V モデルを、オブジェクトの存在、およびテキストプロンプトにおけるオブジェクトの色と数の正しさについて評価する。具体的には、Detection-Score、Count-Score、Color-Scoreを以下のように評価する。

1. Detection-Score($S_{Det}$)：ビデオ全体の平均的なオブジェクトの存在感を測定し、

$$S_{Det} = \frac{1}{M_1}\sum_{i=1}^{M_1}\left(\frac{1}{N}\sum_{t=1}^{N}\sigma_t^i\right) \tag{4}$$

   として計算される。ここで、$M_1$ はオブジェクトを含むプロンプトの数であり、$\sigma_t^i$ はビデオ $i$ のフレーム $t$ の検出結果（オブジェクトが検出された場合は1、そうでない場合は0）である。
2. Count-Score($S_{Count}$)：

$$S_{Count} = \frac{1}{M_2}\sum_{i=1}^{M_2}\left(1 - \frac{1}{N}\sum_{t=1}^{N}\frac{\left|c_t^i - \hat{c}^i\right|}{\hat{c}^i}\right) \tag{5}$$

   として計算される平均オブジェクトカウント差を評価する。ここで、$M_2$ はオブジェクトカウントを持つプロンプトの数、$c_t^i$ はビデオ $i$ の検出されたオブジェクトカウントフレーム $t$、$\hat{c}^i$ はビデオ $i$ のグランドトゥルースのオブジェクトカウントである。
3. Color-Score($S_{Color}$)：

$$S_{Color} = \frac{1}{M_3}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{t=1}^{N}s_t^i\right) \tag{6}$$

   として計算される平均色精度を評価する。ここで、$M_3$ はオブジェクトの色を持つプロンプトの数、$s_t^i$ はビデオ $t$ のフレーム $i$ の色精度結果（検出された色が真実の色と一致する場合は1、そうでない場合は0）である。

### 人間分析（Celebrity ID Score）

私たちが収集した実世界のプロンプトに示されているように、生成されたビデオにとって人間は重要である。このため、一般的な顔分析ツールボックスである DeepFace [55] を使用して、人間の顔の正しさも評価した。生成された有名人の顔と、対応する有名人の実画像との距離を計算することによって分析を行う。

$$S_{CIS} = \frac{1}{M_4}\sum_{i=1}^{M_4}\left(\frac{1}{N}\sum_{t=1}^{N}\left(\min_{k\in\{1,\cdots,N_3\}}\mathcal{D}\left(x_t^i, f_k^i\right)\right)\right) \tag{7}$$

ここで $M_4$ は有名人を含むプロンプトの数、$\mathcal{D}(\cdot,\cdot)$ は Deepface の距離関数、$\{f_k^i\}_{k=1}^{N_3}$ はプロンプトiに対して収集された有名人画像、$N_3$ は $3$ に設定される。

### テキスト認識（OCR-Score）

ビジュアル生成のもう一つの難しいケースは、説明文のテキストを生成することである。テキスト生成のための現在のモデルの能力を調べるために、我々は、以前のテキストから画像への評価 [44:2] またはマルチモデルLLM評価法 [10:2] と同様に、光学式文字認識（OCR）モデルのアルゴリズムを利用する。具体的には、PaddleOCRを利用して、各モデルが生成した英文を検出する。次に、単語誤り率（WER）[56]、正規化編集距離（NED）[57]、文字誤り率（CER）[58]を計算し、最後にこれら3つのスコアを平均して OCR-Score を得る。

## 4.3 動きの質

ビデオの場合、動きのクオリティが画像など他の領域との大きな違いだと考えている。このため、私たちは、動きの質を評価システムの主要な評価指標の1つとみなしています。ここでは、以下に紹介する2つの異なる動きの質を考える。

### 行動認識（Action-Score）

人間に関するビデオの場合、事前に訓練されたモデルによって、一般的な行動を簡単に認識することができる。我々の実験では、MMAction2 ツールボックス [59]、特に事前に訓練された VideoMAE V2 [60] モデルを使用して、生成されたビデオ内の人間の行動を推測する。次に、分類精度（グランドトゥルースは入力プロンプトのアクション）を Action-Score とする。この研究では、Kinetics の400アクションクラス [61] に焦点を当てる。これは広く使われており、楽器を演奏するような人間とオブジェクトのインタラクションや、握手やハグを含む人間と人間のインタラクションを包含している。

### 平均フロー（Flow-Score）

また、映像の一般的な動き情報も考慮する。このため、事前に学習させたオプティカル・フロー推定手法である RAFT [62] を用いて、2フレームごとの映像の密なフローを抽出する。次に、これらのフレームの平均フローを計算し、特定の生成ビデオクリップの平均フロースコアを求める。というのも、時間的整合性メトリクスでは識別しにくい静止画を生成する可能性が高い手法もあるからだ。

### 振幅分類スコア（Motion AC-Score）

平均フローに基づき、生成されたビデオの動きの振幅が、テキストプロンプトで指定された振幅と一致しているかどうかをさらに識別する。このため、平均フロー閾値 $\rho$ を設定し、$\rho$ を超えると1つの動画が大きいと見なされるようにし、ここでは主観的な観察に基づいて $\rho$ を2に設定している。生成されたビデオの動きを識別するために、このスコアをマークする。

## 4.4 時間的整合性

時間的な一貫性もまた、私たちが生成した映像において非常に価値のある分野である。そのために、いくつかの計算指標を用いる。以下にそれらを列挙する。

### ワープエラー

まず、ワープエラーを考える。これは、これまでのブラインド時間一貫性手法 [63] [64] [65] で広く使われているものである。詳細には、まず、事前に訓練したオプティカルフロー推定ネットワーク [62:1] を用いて、各2フレームのオプティカルフローを求め、次に、ワープした画像と予測画像のピクセル単位の差分を計算する。2フレームごとにワープの差を計算し、すべてのペアの平均を使って最終的なスコアを算出する。

### 意味的一貫性（CLIP-Temp）

ピクセル単位の誤差の他に、2つのフレーム間の意味的な整合性も考慮する。これは、以前のビデオ編集作品 [8:4] [65:1] でも使用されている。具体的には、生成された映像の2フレームそれぞれについて意味埋め込みを考え、2フレームそれぞれの平均を求めると、次のようになる。

$$S_{CT} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N-1}\sum_{t=1}^{N-1}\mathcal{C}\left(emb\left(x_t^i\right), emb\left(x_{t+1}^i\right)\right)\right) \tag{8}$$

### 顔の一貫性

CLIP-Temp と同様に、生成された動画の人間の同一性を評価する。具体的には、最初のフレームを参照フレームとして選択し、参照フレームの埋め込みと他のフレームの埋め込みとの余弦類似度を計算する。そして、その類似度を平均して最終スコアとする。

$$S_{FC} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N-1}\sum_{t=1}^{N-1}\mathcal{C}\left(emb\left(x_{t+1}^{i}\right), emb\left(x_{1}^{i}\right)\right)\right) \tag{9}$$

## 4.5 ユーザーの意見

上記の客観的な指標に加えて、私たちはユーザーの意見を聞くために、主要な5つの側面についてユーザー調査を実施している。これらの側面には以下が含まれる。

1. **映像の品質**。これは、生成されたビデオの品質を示すもので、スコアが高いほど、ぼやけやノイズなどの映像劣化がないことを示す。
2. **テキストとビデオの位置合わせ**。この意見は、生成されたビデオと入力テキスト-プロンプトの間の関係を考慮し、生成されたビデオは、間違ったカウント、属性、および関係を持つ低品質のサンプルとみなされる。
3. **動きの品質**。この指標では、ユーザーはビデオから生成されたモーションの正しさを識別する必要がある。
4. **時間的整合性**。時間的一貫性と動きの質は異なる。動きの質では、ユーザーは質の高い動きにランクをつける必要がある。しかし、時間的整合性においては、各映像のフレーム単位の整合性だけを考慮すればよい。
5. **主観的な類似性**。この指標は美的指標に似ており、値が高いほど、生成された動画が一般的に人間の好みを満たしていることを示す。

| Dementions | Metrics | ModelScope-XL [29:1] | ZeroScope [28:1] | Floor33 [20:3] | PikaLab [9:4] | Gen2 [8:5] |
|---|---|---|---|---|---|---|
| **Video Quality** | $\text{VQA}_A \uparrow$ | 97.72 | 95.95 | 98.11 | **99.32** | 99.04 |
| | $\text{VQA}_T \uparrow$ | 6.09 | 6.50 | 7.60 | 8.69 | **10.13** |
| | IS ↑ | **15.99** | 13.35 | 15.10 | 13.66 | 12.57 |
| **Text-video Alignment** | CLIP-Score ↑ | 20.62 | 20.20 | **21.15** | 20.72 | 20.90 |
| | BLIP-BLUE ↑ | 22.42 | 21.20 | **23.67** | 21.89 | 22.33 |
| | SD-Score ↑ | 68.50 | 67.79 | 69.04 | 69.14 | **69.31** |
| | Detection-Score ↑ | 49.59 | 45.80 | **55.00** | 50.49 | 52.44 |
| | Color-Score ↑ | 40.10 | **46.35** | 35.07 | 36.57 | 32.29 |
| | Count-Score ↑ | 47.67 | 47.88 | **57.63** | 56.46 | 57.19 |
| | OCR Score ↓ | 83.74 | 82.58 | **81.09** | 81.33 | 92.94 |
| | Celebrity ID Score ↑ | 45.66 | **45.96** | 45.24 | 43.43 | 44.58 |
| **Motion Quality** | Action Score ↑ | 73.75 | 71.74 | **74.48** | 69.84 | 54.99 |
| | Motion AC-Score → | 26.67 | 53.33 | 60.00 | 40.00 | 40.00 |
| | Flow-Score → | 2.28 | 1.66 | 2.23 | 0.11 | 0.18 |
| **Temporal Consistency** | CLIP-Temp ↑ | 99.72 | 99.84 | 99.58 | **99.97** | 99.92 |
| | Warping Error ↓ | 73.04 | 80.32 | 69.77 | 66.88 | **58.19** |
| | Face Consistency ↑ | 98.89 | 99.33 | 99.05 | **99.64** | 99.06 |

**表2**：映像の質、テキストと映像の整合性、動きの質、時間的整合性の観点から見た生の結果。

評価のために、ModelScope [29:2]、ZeroScope [28:2]、Gen2 [8:6]、Floor33 [66]、PikaLab [9:5] の5つの最先端手法で、提供されたプロンプトベンチマークを使ってビデオを生成し、合計 2,500 本のビデオを得た。公平に比較するために、Gen2 と PikaLab のアスペクト比を 16:9 に変更し、他の方法と比較する。また、PikaLab は視覚的透かしなしでコンテンツを生成することができないため、公正な比較のために、PikaLab の透かしを他のすべての方法に追加した。また、プロンプトをよく理解できないユーザーがいることも考慮している。この目的のために、SDXL [5:3] を使用して、各プロンプトの3つの参照画像を生成し、ユーザの理解を助ける。これはまた、モデルのテキスト-ビデオアラインメントを評

価するための SD-Score を設計するきっかけとなる。それぞれの指標について、3人のユーザーに 1〜5 の間で意見を求め、値が大きいほど良いアライメントであることを示す。平均スコアを最終的なラベリングとして使用し、$[0, 1]$ の範囲に正規化する。

ユーザーデータを収集した後、T2V アルゴリズムのより信頼性の高いロバストな評価を確立することを目的として、評価指標のヒューマンアライメントを実施する。最初に、我々は、特定の側面におけるユーザーの意見に対する人間のスコアを近似するために、上記の個々のメトリックを使用してデータのアライメントを行う。我々は、自然言語処理の評価 [67] [68] の研究にヒントを得て、各次元のパラメータを適合させるために線形回帰モデルを採用する。具体的には、4つの異なる手法から無作為に300サンプルをフィッティングサンプルとして選び、残りの200サンプルは提案手法の有効性を検証するために残した（表4の通り）。係数パラメータは、人間のラベルと線形回帰モデルからの予測値との間の残差二乗和を最小化することによって得られる。次の段階では、これら4つの側面の整合結果を統合し、平均スコアを算出して、T2Vアルゴリズムの性能を効果的に表す包括的な最終スコアを得る。このアプローチは、評価プロセスを合理化し、モデルの性能を明確に示す。

| Aspects | Methods | Spearman's$\rho$ | Kendall's$\phi$ |
|---|---|---|---|
| **Visual Quality** | $\text{VQA}_A$ | 42.1 | 30.5 |
| | $\text{VQA}_T$ | 49.3 | 35.9 |
| | Avg. | 45.9 | 33.7 |
| | **Ours** | **50.2** | **37.6** |
| **Motion Amplitude** | Motion AC | −16.9 | −13.1 |
| | Flow-Score | −32.9 | −23.1 |
| | Avg. | −27.8 | −20.4 |
| | **Ours** | **32.1** | **24.0** |
| **Temporal Consistency** | CLIP-Temp. | 50.0 | 35.8 |
| | Warp Error | 36.1 | 27.1 |
| | Avg. | 37.2 | 27.9 |
| | **Ours** | **50.0** | **36.0** |
| **TV Alignment** | SD-Score | 10.0 | 6.9 |
| | CLIP-Score | 14.4 | 10.1 |
| | Avg. | 20.2 | 14.0 |
| | **Ours** | **30.5** | **21.7** |

**表4**：**訂正分析**。テキストからビデオへの変換における、いくつかの客観的指標と人間の判断との相関。相関計算にはスピアズマンの $\rho$ とケンドールの $\phi$ を使用。

# 5 結果

ベンチマークプロンプトの中から500個のプロンプトについて評価を行い、各プロンプトには評価用の回答として追加情報のメタファイルが用意されている。ModelScope [29:3]、Floor33 Pictures [66:1]、ZeroScope [28:3] など、利用可能なすべての高解像度T2Vモデルを用いて動画を生成する。分類子を使わないガイダンスなど、すべてのハイパーパラメータはデフォルト値のままにしておく。サービスベースのモデルについては、代表的な作品であるGen2 [8:7] と PikaLab [9:6] のパフォーマンスを評価した。少なくとも 512p の高画質ウォーターマークフリービデオを生成する。評価の前に、これらのモデルの能力、生成された解像度、fps など、各ビデオタイプの違いを表1に示す。速度の比較については、NVIDIA A100で利用可能なすべてのモデルを実行した。利用できないモデルについては、そのモデルをオンラインで実行し、おおよその時間を計測する。PikaLab と Gen2 は、追加のハイパーパラメータによってモーションとカメラを制御する能力も持っていることに注意してほしい。また、調整可能なパラメーターは多いが、比較的公平な比較のためにデフォルト設定のままにしている。

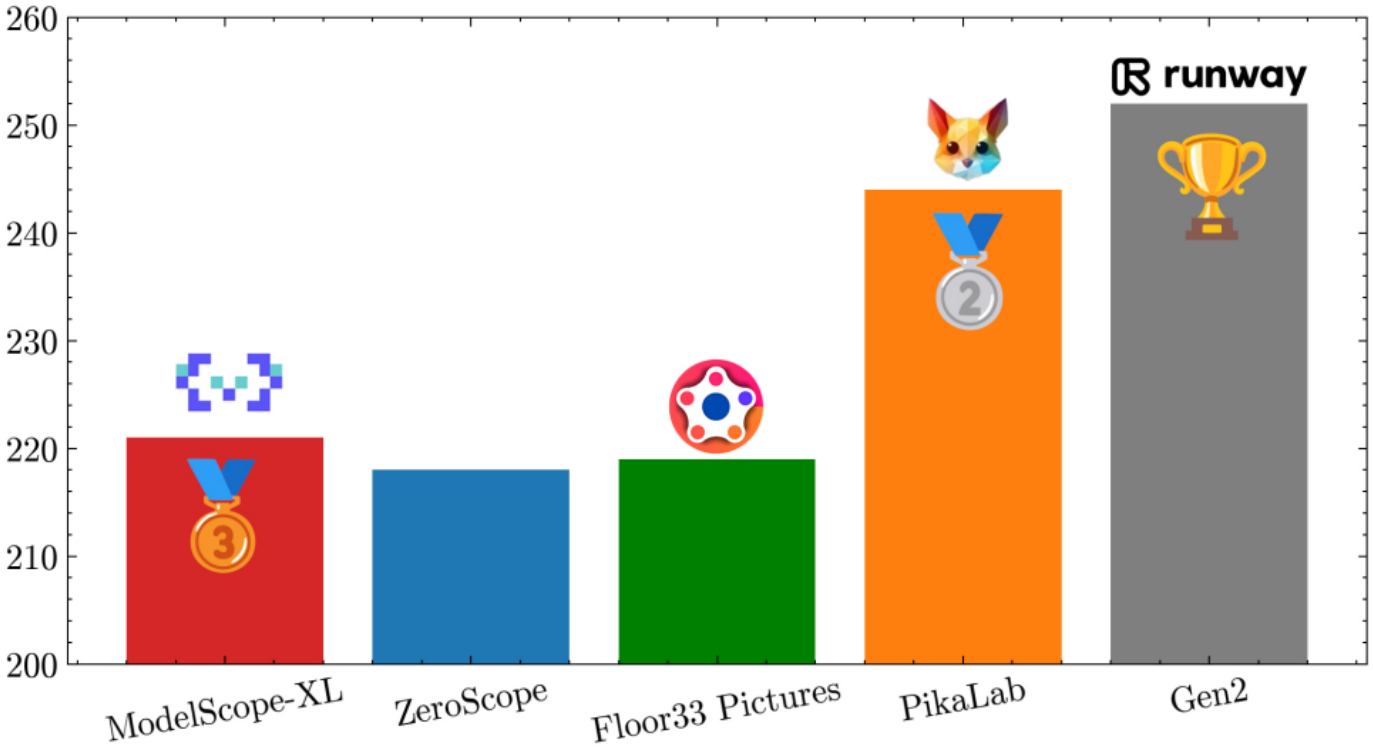



**図5**：**EvalCrafterベンチマークでの総合比較結果**。

| | Visual Quality | Text-Video Alignment | Motion Quality | Temporal Consistency |
|---|---|---|---|---|
| **ModelScope-XL** | 55.23(5) | 47.22(4) | 59.41(2) | 59.31(4) |
| **ZeroScope** | 56.37(4) | 46.18(5) | 54.26(4) | 61.19(3) |
| **Floor33 Pictures** | 59.53(3) | 51.29(3) | 51.97(5) | 56.36(5) |
| **PikaLab** | 63.52(2) | **54.11(1)** | 57.74(3) | 69.35(2) |
| **Gen2** | **67.35(1)** | 52.30(2) | **62.53(1)** | **69.71(1)** |

**表3**：4つの異なる側面から、人間による優先順位を揃えた結果（カッコ内は各側面の順位）。

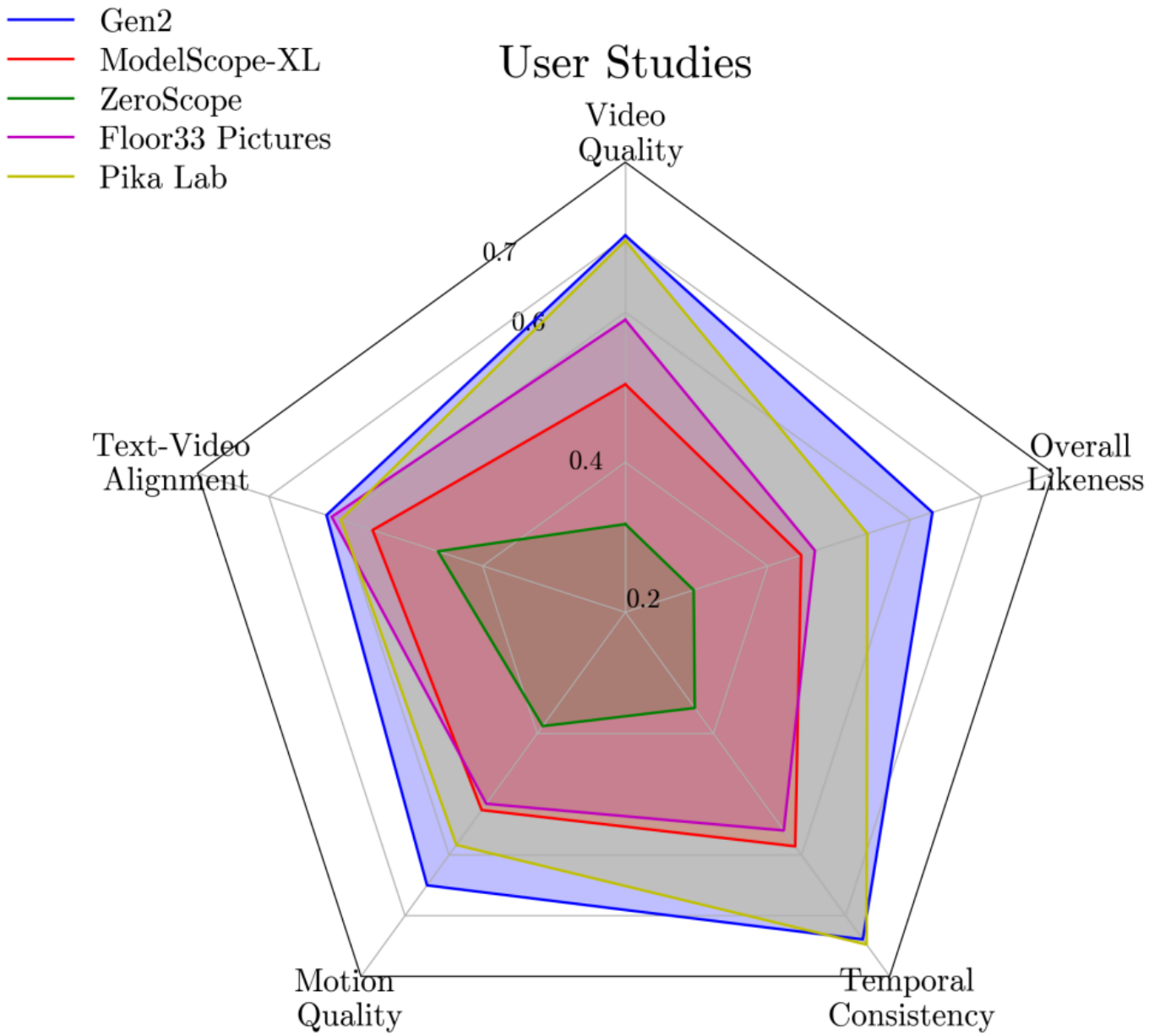



**図6**：**ユーザー調査による生の評価。**

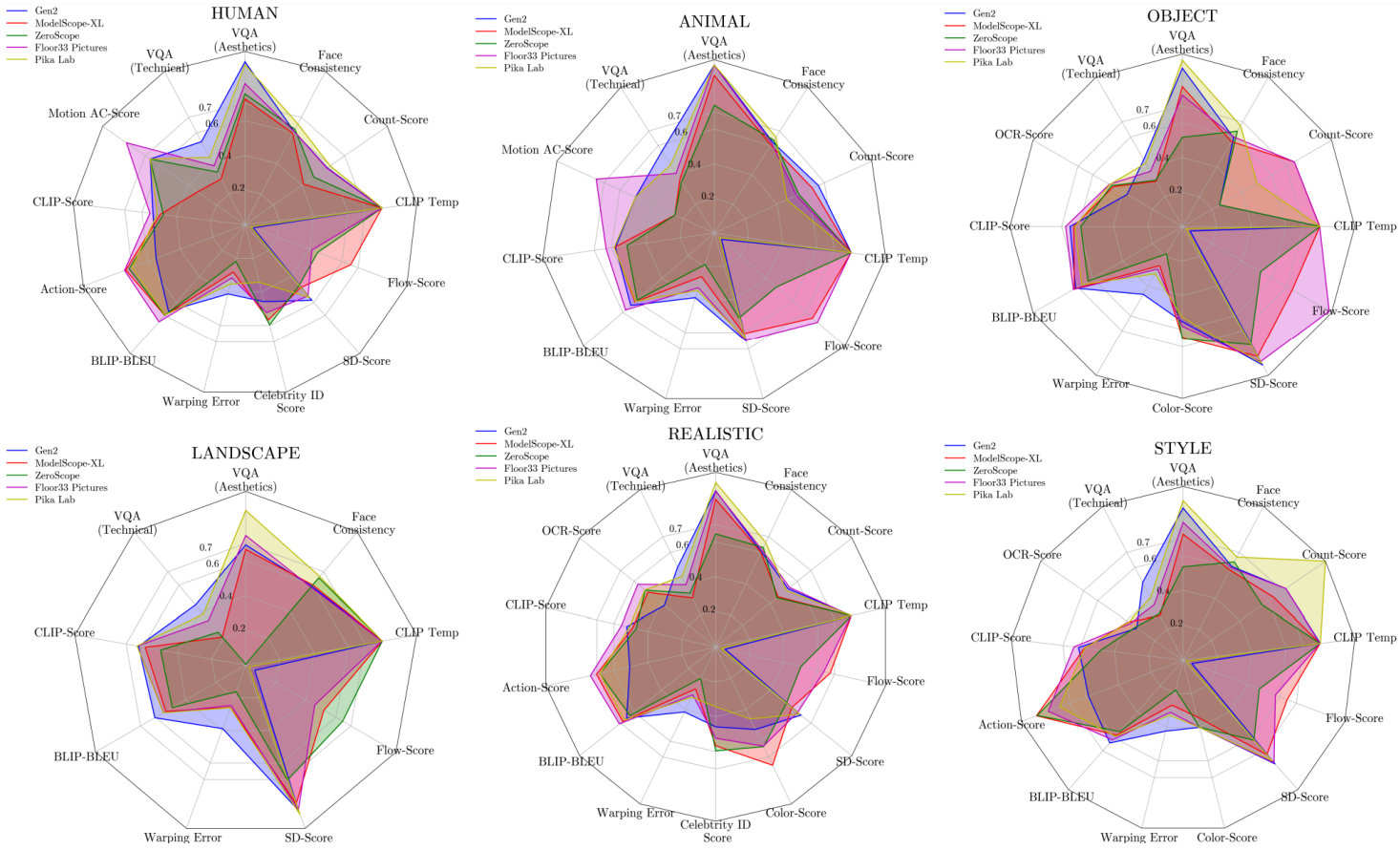



**図7：さまざまな側面における生の結果。**生成されたビデオのメタタイプの性能を評価するために、4つの主要なメタタイプ（動物、人間、風景、物体）を考慮する。各タイプには、きめ細かい属性ラベルを持つ複数のプロンプトが含まれる。それぞれのプロンプトに対して、ビデオのスタイルも考慮するが、さらに多様なプロンプトがある。上の現実的でスタイリッシュな図のように。(メトリクスの値は、より見やすくするために正規化されている。ワーピング・エラーとOCRスコアを前処理しているため、この図ではこの2つのメトリクスの値が大きいほど性能が良いことを示している)

まず、図5に全体的なヒューマンアライメントの結果を示し、表3にベンチマークのさまざまな側面を示す。 これにより、ベンチマークの最終的な主要指標が得られる。最後に、図7と同様に、我々のベンチマークにおける4つの異なるメタタイプ（すなわち、動物、人間、風景、物体）と、我々のベンチマークにおける2つの異なるタイプの動画（すなわち、一般、スタイル）に対する各手法の結果を示す。各手法の客観的・主観的指標を比較するために、各指標の生データを表1と図6に示す。5.1節で詳細な分析を行う。

## 5.1 分析

### 発見① 単一の指標でモデルを評価するのは不公平である

表3から、モデルの順位はこれらの側面で大きく異なっており、パフォーマンスを包括的に理解するためには、多面的な評価アプローチが重要であることが浮き彫りになった。例えば、Gen2 が Visual Quality、Motion Quality、Temporal Consistency の点で他のモデルを凌駕する一方で、PikaLab は Text-Video Alignment で優れたパフォーマンスを示している。

### 発見② メタタイプ別にモデルの能力を評価する必要がある

図7に示すように、ほとんどのメソッドが、メタタイプによって大きく異なる値を示している。例えば、Gen2 [8:8] は、我々の実験では T2V の全体的なアライメントが最も優れているが、この方法から生成された動画は、行動認識モデルでは認識しにくい。我々は主観的に、Gen2 は主にテキストプロンプトからのクローズアップショットを弱い動きの振幅で生成する。

### 発見③ ユーザーは見た目の品質よりも、T2Vのアライメントの悪さに寛容である

図7と表2に示すように、Gen2 [8:9] は、すべてのテキストとビデオのアライメント・メトリクスで良い結果を出すことはできないが、時間的な整合性、視覚的な品質、小さな動きの振幅が良いため、ユーザはほとんどの場合、このモデルの結果を許容する。

### 発見④ テキストプロンプトから直接カメラの動きを制御できない

いくつかの追加のハイパーパラメータは、追加のコントロールハンドルとして設定することができるが、現在の T2V のテキストエンコーダは、カメラの動きのようなオープンワールドのプロンプトの背後にある理由の理解がまだ不足している。

### 発見⑤ 視覚に訴えることは、生成された解像度と正の相関はない

表1に示すように、gen2 [8:10] は解像度が最も小さいが、表2、図6に示すように、人間と客観的メトリクスの両方が、この方法は最高の視覚的品質を持ち、不自然さが少ないとみなしている。

### 発見⑥ 動きの振幅が大きいからといって、ユーザーにとってより良いモデルであるとは限らない

図6から、PikaLab [9:7] と Gen2 [8:11] の2つの小さなモーションモデルは、Floor33 Pictures [66:2] の大きなモーションモデルよりも、ユーザの選択において良いスコアを獲得している。下手で理不尽な動きの映像よりも、わずかな動きの映像の方が、ユーザーは見やすいのだ。

### 発見⑦ テキスト記述からテキストを生成するのはまだ難しい

これらのモデルのOCRスコアを報告するが、テキストプロンプトから現実的なフォントを生成するのはまだ難しい。テキストプロンプトから高品質で一貫性のあるテキストを生成するには、ほぼすべての方法が公平である。

### 発見⑧ 現在のビデオ生成モデルは、依然として一発で結果を生成している

表2にあるように、どの手法も CLIP-Temp の一貫性が非常に高いことを示している。これは、各フレームがフレーム間で非常に類似した意味を持っていることを意味する。そのため、現在の T2V モデルは、複数のトランジションやアクションを含む長い動画以外のシネマグラフを生成する可能性が高い。

### 発見⑨ 最も価値ある客観的指標

客観的な指標を実際のユーザーに合わせることで、一つの側面から価値ある指標を見出すこともできる。例えば、表2と表3によると、SD-Score と CLIP-Score はどちらもテキストとビデオのアライメントに価値がある。VQAT と VQAA は、視覚的な品質評価にも価値がある。

### 発見⑩ Gen2 も完璧ではない

Gen2 [8:12] は我々の評価で総合トップのパフォーマンスを達成したが、まだ複数の問題を抱えている。例えば、Gen2 はプロンプトから複雑なシーンのビデオを生成するのが難しい。Gen2 は、表1の IS メトリック（ネットワークでも識別されにくい）にも反映されているように、人間にも動物にも奇妙なアイデンティティを持つが、他の手法にはそのような問題はない。

### 発見⑪ オープンソースとクローズドソースの T2V モデルの間には大きな性能差が存在する

表3を参照すると、ModelScope-XL や ZeroScope のようなオープンソースのモデルは、PikaLab [9:8] や Gen2 [8:13] のようなクローズドソースのモデルと比較して、ほとんどすべての面でスコアが低いことがわかる。このことは、オープンソースの T2V モデルが、クローズドソースの T2V モデルの性能レベルに達するには、まだ改善の余地があることを示している。

## 5.2 人間の嗜好アライメントに関するアブレーション

人間のスコアとのアライメントにおける我々のモデルの有効性を示すために、スピアマンの順位相関係数 [69] とケンドールの順位相関係数 [70] を計算した。これらの係数は、表4に記載されているように、我々のメソッドの結果と人間のスコアとの間の関連性の強さと方向性についての洞察を与えてくれる。この表から、提案された重み付け方法は、直接平均化するよりも、未見の200サンプルでより良い相関を示している（最初に $[0, 1]$ の範囲になるように、すべてのデータを100で割る）。もうひとつの興味深い発見は、現在のモーションアンプリチュードのスコアはすべて、ユーザーの選択とは関係がないということだ。私たちは、人間は振幅よりも運動の安定性を重視すると主張する。しかし、私たちのフィッティング方法は、より高い相関を示している。

## 5.3 制限

T2V 生成の評価はすでに一歩前進しているが、まだ課題は多い。

1. 現在、私たちは500のプロンプトをベンチマークとして収集しているが、実際の状況は非常に複雑である。より多くのプロンプトは、より詳細なベンチマークを示す。

2. 一般的な感覚の動きの良さを評価するのも難しい。しかし、マルチモデル LLM や大規模なビデオ基礎モデルの時代には、より優れたより大規模なビデオ理解モデルがリリースされ、それらを私たちの指標として使用することができると信じている。
3. アライメントに使用されたラベルは、3人の人間の注釈者のみから収集されたものであるため、結果に多少のバイアスが生じる可能性がある。この限界に対処するため、より正確で偏りのない評価を確実にするために、アノテーターのプールを拡大し、より多様なスコアを集める予定である。

# 6 結論

オープンワールドの大規模なジェネレーティブ・モデルの能力をさらに発見することは、より良いモデル設計と活用のために不可欠である。本論文では、大規模かつ高品質な T2V モデルの評価の第一歩を踏み出す。この目標を達成するために、まず T2V 評価のための詳細なプロンプトベンチマークを構築した。一方、T2V モデルの性能を評価するために、映像品質、テキストと映像のアライメント、対象物、動きの品質について、いくつかの客観的な評価指標を与える。最後に、ユーザ調査を実施し、ユーザスコアと客観的メトリクスをマッチングさせる新しいアラインメント手法を提案する。実験では、提案手法の能力がユーザの意見をうまく調整できることを示し、T2V手法の正確な評価指標を与えた。

# 参考


1. Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717, 2018. ↩ ↩ ↩ ↩
2. Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. Advances in neural information processing systems, 29, 2016. ↩ ↩ ↩
3. OpenAI. Gpt-4 technical report, 2023. ↩ ↩ ↩ ↩ ↩ ↩
4. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models, 2021. ↩ ↩ ↩
5. Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023. ↩ ↩ ↩ ↩
6. Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303, 2022. ↩ ↩ ↩
7. Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792, 2022. ↩ ↩ ↩
8. Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. arXiv preprint arXiv:2302.03011, 2023. ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩
9. Pika Lab discord server. https://www.pika.art/. Accessed: 2023-08-30. ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩
10. Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. Jul 2023. ↩ ↩ ↩
11. Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897, 2023. ↩ ↩
12. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013. ↩
13. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014. ↩ ↩
14. Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation, 2022. ↩
15. Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3626–3636, 2022. ↩
16. Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In International Conference on Learning Representations, 2022. ↩
17. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020. ↩
18. Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 2 ↩

19. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021. ↩ ↩ ↩
20. Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022. ↩ ↩ ↩ ↩
21. Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023. ↩ ↩ ↩
22. Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018, 2022. ↩ ↩
23. Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725, 2023. ↩
24. David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation, 2023. ↩
25. Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models, 2023. ↩
26. Morph studio discord server. https://www.morphstudio.com/. Accessed: 2023-08-30. ↩
27. Fulljourney discord server. https://www.fulljourney.ai/. Accessed: 2023-08-30. ↩ ↩
28. Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w. Accessed:2023-08-30. ↩ ↩ ↩ ↩
29. Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571, 2023. ↩ ↩ ↩ ↩
30. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. ↩
31. Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023. ↩
32. Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. arXiv preprint arXiv:2305.14938, 2023. ↩
33. Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874, 2021. ↩
34. Qiantong Xu, Fenglu Hong, Bo Li, Changran Hu, Zhengyu Chen, and Jian Zhang. On the tool manipulation capability of open-source large language models, 2023. ↩
35. Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. ↩
36. Gu Zhouhong, Zhu Xiaoxuan, Ye Haoning, Zhang Lin, Wang Jianchen, Jiang Sihang, Xiong Zhuozhi, Li Zihan, He Qianyu, Xu Rui, Huang Wenhao, Zheng Weiguo, Feng Hongwei, and Xiao Yanghua. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. arXiv:2304.11679, 2023. ↩
37. Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109, 2023. ↩
38. Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023. ↩
39. Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023, 2023. ↩
40. Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178, 2023. ↩
41. Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 35:36479–36494, 2022. ↩ ↩
42. Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. ↩ ↩
43. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. Endto-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020. ↩

44. EslamMohamed Bakr, Pengzhan Sun, Xiaoqian Shen, FaizanFarooq Khan, LiErran Li, and Mohamed Elhoseiny. Hrsbench: Holistic, reliable and scalable benchmark for text-toimage models. Apr 2023. ↩ ↩ ↩
45. George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41, 1995. ↩
46. Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0. ↩
47. Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In International Conference on Computer Vision (ICCV), 2023. ↩ ↩
48. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009. ↩
49. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015. ↩
50. Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023.\ ↩
51. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. ↩
52. Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. arXiv preprint arXiv:2305.06558, 2023. ↩
53. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023. ↩
54. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. ↩
55. Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In 2021 International Conference on Engineering and Emerging Technologies (ICEET), pages 1–4. IEEE, 2021. ↩
56. Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. Speech Communication, 38(1-2):19–28, 2002. ↩
57. Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on largescale street view text with partial labeling-rrc-lsvt. In 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1557–1562. IEEE, 2019. ↩
58. Andrew Cameron Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In Eighth International Conference on Spoken Language Processing, 2004. ↩
59. MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. ↩
60. Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking, 2023. ↩
61. Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017. ↩
62. Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pages 402–419. Springer, 2020. ↩ ↩
63. Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In Proceedings of the European conference on computer vision (ECCV), pages 170–185, 2018. ↩
64. Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. In Advances in Neural Information Processing Systems, 2020. ↩
65. Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv:2303.09535, 2023. ↩ ↩
66. Floor33 pictures discord server. https://www.morphstudio.com/. Accessed: 2023-08-30. ↩ ↩ ↩
67. Kallirroi Georgila, Carla Gordon, Volodymyr Yanov, and David Traum. Predicting ratings of real dialogue participants from artificial data and ratings of human dialogue observers. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 726–734, 2020. ↩
68. Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. International Journal of Computer Vision, 129:1238–1257, 2021. ↩

69. Jerrold H Zar. Spearman rank correlation. Encyclopedia of Biostatistics, 7, 2005. ↩
70. Maurice George Kendall. Rank correlation methods. 1948. ↩