EvalCrafter：大規模ビデオ生成モデルのベンチマークと評価

図1：グレーで示された明確なプロンプトタイプと黒丸に描かれた複数の評価項目を含むようなT2V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 EvalCrafter を提案する。

> We propose EvalCrafter, a comprehensive framework for benchmarking and evaluating the text-to-video models, including the well-defined prompt types in grey and the multiple evaluation aspects in black circles.

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、高画質な動画を生成する様々なオープンソースのモデルや誰でも利用できるサービスが公開されている。しかし、これらの手法では、FVD [1] や IS [2] など、数少ない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、T2V モデルの生成物のための新しいプロンプトリストを作成する。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の客観的指標で評価する。モデルの最終的なリーダーボードを得るために、客観的な指標をユーザーの意見に合わせる。そのために、一連の係数を最適化させた。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。

> The vision and language generative models have been overgrown in recent years. For video generation, various open-sourced models and public-available services are released for generating high-visual quality videos. However, these methods often use a few academic metrics, e.g., FVD [1:1] or IS [2:1], to evaluate the performance. We argue that it is hard to judge the large conditional generative models from the simple metrics since these models are often trained on very large datasets with multi-aspect abilities. Thus, we propose a new framework and pipeline to **exhaustively** evaluate the performance of the generated videos. To achieve this, we first conduct a new prompt list for text-to-video generation by analyzing the real-world prompt list with the help of the large language model. Then, we evaluate the state-of-the-art video generative models on our carefully designed benchmarks, in terms of visual qualities, content qualities, motion qualities, and text-caption alignment with around 18 objective metrics. To obtain the final leaderboard of the models, we also fit a series of **coefficients** to align the objective metrics to the users' opinions. Based on the proposed opinion alignment method, our final score shows a higher correlation than simply averaging the metrics, showing the effectiveness of the proposed evaluation method.

1 はじめに

大規模生成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 もとい 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、これら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。

> The charm of the large generative models is sweeping the world. e.g., the well-known ChatGPT and GPT4 [3:1] have shown human-level abilities in several aspects, including coding, solving math problems, and even visual understanding, which can be used to interact with our human beings using any knowledge in a conversational way. As for the generative models for visual content creation, Stable Diffusion [4:1] and SDXL [5:1] play very important roles since they are the most powerful publicly available models that can generate high-quality images from any text prompts.

T2I 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。初期のモデル（ImagenViedo [6]、Make-A-Video [7] ）は、カスケードされた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 [8] と MagicVideo [9] 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 [10] や PikaLabs [11] 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 [1:2] のような、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する以前の 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、生成された動画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 [3:2]、MLLM [12]、text-to-image [13] 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、T2I や対話の評価とは異なり、これらが無視してきた「動き」と「一貫性」が動画生成にとって非常に重要であることである。

> Beyond text-to-image, taming diffusion model for video generation has also progressed rapidly. Early works (ImagenViedo [6:1], Make-A-Video [7:1]) utilize the cascaded models for video generation directly. Powered by the image generation priors in Stable Diffusion, LVDM [8:1] and MagicVideo [9:1] have been proposed to train the temporal layers to efficiently generate videos. Apart from the academic papers, several commercial services also can generate videos from text or images. e.g., Gen2 [10:1] and PikaLabs [11:1]. Although we can not get the technique details of these services, they are not evaluated and compared with other methods. However, all current large text-to-video (T2V) model only uses previous GAN-based metrics, like FVD [1:3], for evaluation, which only **concerns** the distribution matching between the generated video and the real videos, other than the pairs between the text prompt and the generated video. Differently, we argue that a good evaluation method should consider the metrics in different aspects, e.g., the motion quality and the temporal consistency. Also, similar to the large language models, some models are not publicly available and we can only get access to the generated videos, which further increases the difficulties in evaluation. Although the evaluation has progressed rapidly in the large generative models, including the areas of LLM [3:3], MLLM [12:1], and textto-image [13:1], it is still hard to directly use these methods for video generation. The main problem here is that different from text-to-image or dialogue evaluation, motion and consistency are very important to video generation which previous works ignore.

我々は、動画のための大規模マルチモーダル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 [3:4] 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 T2I 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。これらの意見を得た後、各目的回帰モデルの係数を訓練し、評価スコアをユーザーの選択に合わせることで、モデルの最終スコアを得ることができ、また訓練された係数を用いて新しい動画を評価することができる。

> We make the very first step to evaluate the large multimodality generative models for video. In detail, we first build a comprehensive prompt list containing various everyday objects, attributes, and motions. To achieve a balanced distribution of well-known concepts, we start from the welldefined meta-types of the real-world knowledge and utilize the knowledge of large language models, e.g., ChatGPT [3:5], to extend our meta-prompt to a wide range. Besides the prompts generated by the model, we also select the prompts from real-world users and text-to-image prompts. After that, we also obtain the metadata (e.g., color, size, etc.) from the prompt for further evaluation usage. Second, we evaluate the performance of these larger T2V models from different aspects, including the video visual qualities, the text-video alignment, and the motion quality and temporal consistency. For each aspect, we use one or more objective metrics as the evaluation metrics. Since these metrics only reflect one of the abilities of the model, we also conduct a multi-aspects user study to judge the model in terms of its

> qualities. After obtaining these opinions, we train the coefficients of each objective regression model to align the evaluation scores to the user's choice, so that we can obtain the final scores of the models and also evaluate the new video using the trained coefficients.

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、生成動画の評価のために、動画の視覚的品質、動画の動きの品質、およびテキストと動画への変換の側面を考慮する。各アスペクトについて、人間の意見を照合し、また人間による照合によって提案メトリックの有効性を検証する。
- T2V モデルのさらなる学習に役立つ可能性のある幾つかの結論と発見についても議論した。

> Overall, we summarize the contribution of our paper as:
>
> ・We make the first step of evaluating the large T2V model and build a comprehensive prompt list with detailed annotations for T2V evaluation.
>
> ・We consider the aspects of the video visual quality, video motion quality, and text-video alignment for the evaluation of video generation. For each aspect, we align the opinions of humans and also verify the effectiveness of the proposed metric by human alignment.
>
> ・During the evaluation, we also discuss several conclusions and findings, which might be also useful for further training of the T2V generation models.

2 関連研究

2.1 Text-to-Video 生成と評価

T2V 生成は、与えられたテキストプロンプトから動画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 [14]）または生成的敵対ネットワーク（GAN [15]）によってビデオを生成していた。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 [16] や風景 [17] [18] 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 [19]、ビデオ拡散モデル [20]、大規模 T2V 事前学習 [21] の急速な発展に伴い、現在の方法では、生成前に、より強力な T2V 事前学習モデルを利用する。例えば、Make-A-Video [7:2] や Imagen-Video [6:2] 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動画を生成する。LVDM [8:2]、Align Your latent [22]、MagicVideo [9:2]は、時間的なアテンション層やトランスフォーマー層を追加することで、潜在的 T2V モデルを動画領域に拡張している。AnimateDiff [23] は、パーソナライズされた T2V 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 [10:2] や Gen2 [10:3] などのオンラインモデルサービスでは、完全な T2V 生成や条件付き動画生成において、高品質な動画生成能力を示している。Discord ベースのサーバーでは、Pika-Lab [11:2]、Morph Studio [26]、FullJourney [27]、Floor33 Pictures [8:3] 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 T2V（または I2V）モデルもある。

> T2V generation aims to generate videos from the given text prompts. Early works generate the videos through Variational AutoEncoders (VAEs [14:1]) or generative adversarial network (GAN [15:1]). However, the quality of the generated videos is often low quality or can only work on a specific domain, e.g., face [16:1] or landscape [17:1] [18:1]. With the rapid development of the diffusion model [19:1], video diffusion model [20:1], and large-scale text-image pretraining [21:1], current methods **utilize** the stronger text-to-image pre-trained model prior to generation. e.g., Make-A-Video [7:3] and Imagen-Video [6:3] train a cascaded video diffusion model to generate the video in several steps. LVDM [8:4], Align Your latent [22:1] and MagicVideo [9:3] extend the latent text-to-image model to video domains by adding additional **temporal attention** or transformer layer. AnimateDiff [23:1] shows a good visual quality by utilizing the personalized text-toimage model. Similar methods are also been proposed by SHOW-1 [24:1] and LAVIE [25:1] T2V generation also raises the **enthusiasm** of **commerce** or **non-commerce** companies. For online model services, e.g., Gen1 [10:4] and Gen2 [10:5], show the abilities of the high-quality generated video in the fully T2V generation or the conditional video generation. For discord-based servers, Pika-Lab [11:3], Morph Studio [26:1], FullJourney [27:1] and Floor33 Pictures [8:5] also show very **competitive** results. Besides, there are also some popular open-sourced text (or image)-to-video models, e.g., ZeroScope [28:1], ModelScope [29:1].

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 [1:4]（LVDM [8:6]、MagicVideo [9:4]、Align Your Latent [22:2]）、IS [2:2]（Align Your Latent [22:3]）、CLIP 類似性 [21:2]（Gen1 [10:6]、Imagen Video [6:4]、Make-A-Video [7:4]）、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 T2V 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。

> However, these methods still **lack** a fair and detailed benchmark to evaluate the advantages of each method. For example, they only evaluate the performance using FVD [1:5] (LVDM [8:7], MagicVideo [9:5], Align Your Latent [22:4]), IS [2:3] (Align Your Latent [22:5]), CLIP similarity [21:3] (Gen1 [10:7], Imagen Video [6:5], Make-A-Video [7:5]), or user studies to show the performance level. These metrics might only perform well on previous in-domain text-to-image generation methods but ignore the alignment of input text, the motion quality, and the temporal consistency, which are also important for T2V generation.

2.2 大規模生成モデルの評価

大規模な生成モデル [3:6] [5:2] [4:2] [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指標を設計する。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 [12:2] 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。

> Evaluating the large generative models [3:7] [5:3] [4:3] [30:1] [31:1] is a big challenge for both the NLP and vision tasks. For the large language models, current methods design several metrics in terms of different abilities, question types, and user platform [32:1] [33:1] [34:1] [35:1] [36:1]. More details of LLM evaluation and Multi-model LLM evaluation can be found in recent **surveys** [37:1] [39:1]. Similarly, the evaluation of the multimodal generative model also draws the attention of the researchers [39:2] [40:1]. For example, Seed-Bench [12:3] generates the VQA for multi-modal large language model evaluation.

ビジュアル生成タスクのモデルについて、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、T2I 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 [3:8] を用いてプロンプトを生成し、T2I モデルの13のスキルを評価するために17の指標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 [13:2] 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。

> For the models in visual generation tasks, Imagen [41:1] only evaluates the model via user studies. DALL-Eval [42:1] assesses the visual reasoning skills and social basis of the text-to-image model via both user and object detection algorithm [43:1]. HRS-Bench [44:1] proposes a **holistic** and **reliable** benchmark by generating the prompt with ChatGPT [3:9] and utilizing 17 metrics to evaluate the 13 skills of the text-to-image model. TIFA [13:3] proposes a benchmark utilizing the visual question answering (VQA). However, these methods still work for text-to-image evaluation or language model evaluation. For T2V evaluation, we consider the quality of motion and temporal consistency.

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。動画生成には時間がかかるため、初期バージョンとして500個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。

> Our benchmark aims to create a trustworthy prompt list to evaluate the abilities of various of T2V models fairly. To achieve this goal, we first collect and analyze the T2V prompt from large-scale real-world users. After that, we propose an automatic pipeline to increase the diversity of the generated prompts so that they can be identified and evaluated by pre-trained computer vision models. Since video generation is time-consuming, we collect 500 prompts as our initial version for evaluation with careful annotation. Below, we give the details of each step.

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 [11:4] の実世界プロンプトの分析。

> The analysis of the real-world prompts from PikaLab Server [11:5].

この問いに答えるために、我々は、FullJourney [27:2] や PikaLab [11:6] 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60万個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、the person、the style、human motion、sceneが支配的で、video、camera、high、quality 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、動画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。

> To answer this question, we collect the prompts from the real-world T2V generation discord users, including the FullJourney [27:3] and PikaLab [11:7]. In total, we get over 600k prompts with corresponding videos and filter them to 200k by removing repeated and meaningless prompts. Our first curiosity is how long a prompt should be generated, as shown in Fig. 2 (a), 90% of the prompts contain the words in the range of $[3, 40]$. We also plot the most important words in Fig. 2 (b) by removing some unclear words like video, camera, high, quality, etc., where the person, the style, human motion, and scene are dominant. **Despite** the above analysis, we also count the word class to decide the meta class of our prompt list. As shown in Fig. 2 (c), we use WordNet [45:1] to identify the meta classes, except for the communication, attribute, and cognition words, the artifacts (human-made objects), human, animal, and the location (landscape) play important roles. We also add the most important word style of Fig. 2 (b) to the metaclass. Overall, we divide the T2V generation into roughly four meta-subject classes, including the human, animal, object, and landscape. For each type, we also consider the motions and styles of each type and the relationship between the current metaclass and other metaclasses to construct the video. Besides, we include the motion which is **relevant** to the main object and important for the video. Finally, we consider the camera motion and the style by template.

3.2 一般的な認識可能プロンプト生成

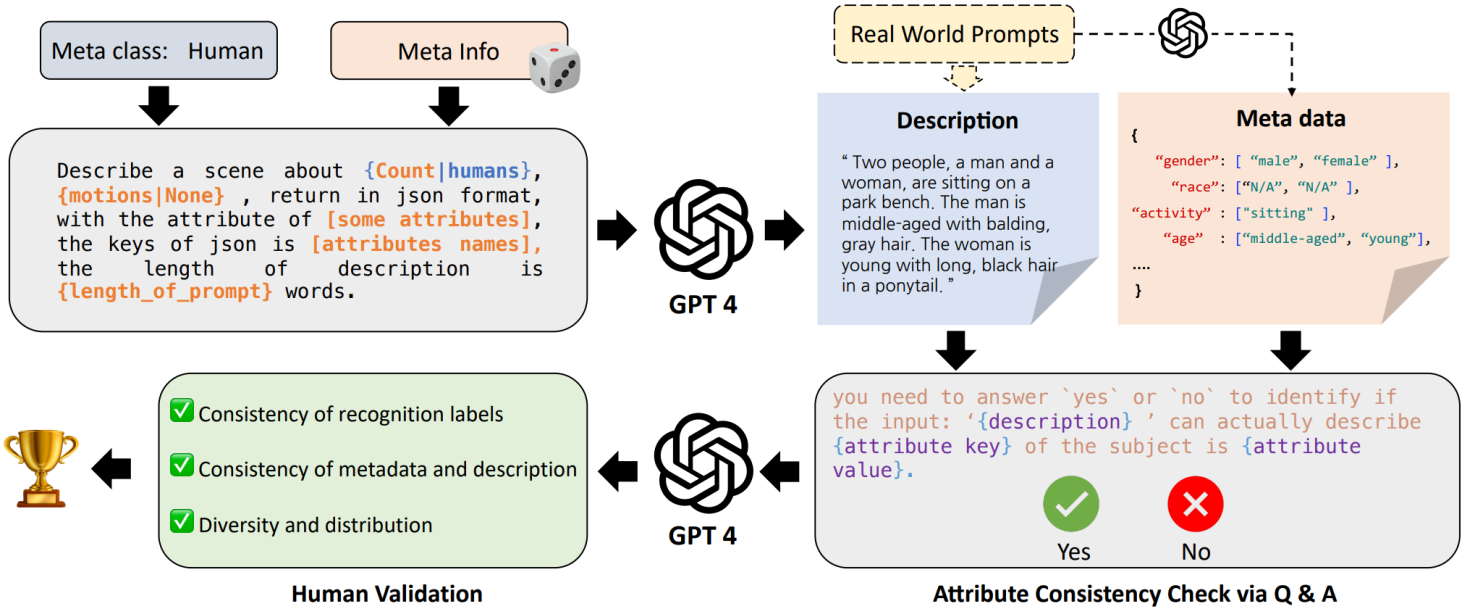

図3：我々は、コンピュータビジョンモデルとユーザーによる T2V 評価のために、詳細なプロンプトを備えた信頼できるベンチマークを生成することを目指している。そのパイプラインを上に示す。

> We aim to generate a trustworthy benchmark with detailed prompts for text-to-video evaluation by computer vision model and users. We show the pipeline above.

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 [3:10] 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 形式で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再度 GPT-4 を使用し、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。

> After deciding the meta classes of our prompt list, we generate the recognizable prompt by the power of a large language model (LLM) and humans. As shown in Fig 3, for each kind of meta class, we let GPT-4 [3:11] describe the scenes about this meta class with randomly sampled meta information along with the attributes of the scenes so that we already know the labels. For example, for humans, we can ask GPT-4 to give us the attributes of humankind, age, gender, clothes, and human activity, which are saved as a JSON file as the ground truth computer vision models. However, we also find that the GPT-4 is not fully perfect for this task, the generated attributes are not very consistent with the generated description. Thus, we **involve** a self-check to the benchmark building, where we also use GPT-4 to identify the similarities of the generated description and the meta data. Finally, we filter the prompts by ourselves to make sure each prompt is correct and meaningful for T2V generation.

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 [42:2] やDraw-Bench [41:2] など、利用可能なT2I 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。

> Since we have already collected a very large scale of prompts from real-world users and there are also available text-to-image evaluation prompts, e.g., DALL-Eval [12] and Draw-Bench [45], we also integrate these prompts to our benchmark list. To achieve this, we first filter and generate the metadata using GPT-4. Then, we choose the suitable prompts with the corresponding metainformation as shown in Fig. 3 and check the consistency of the meta-information.

3.3 ベンチマーク分析

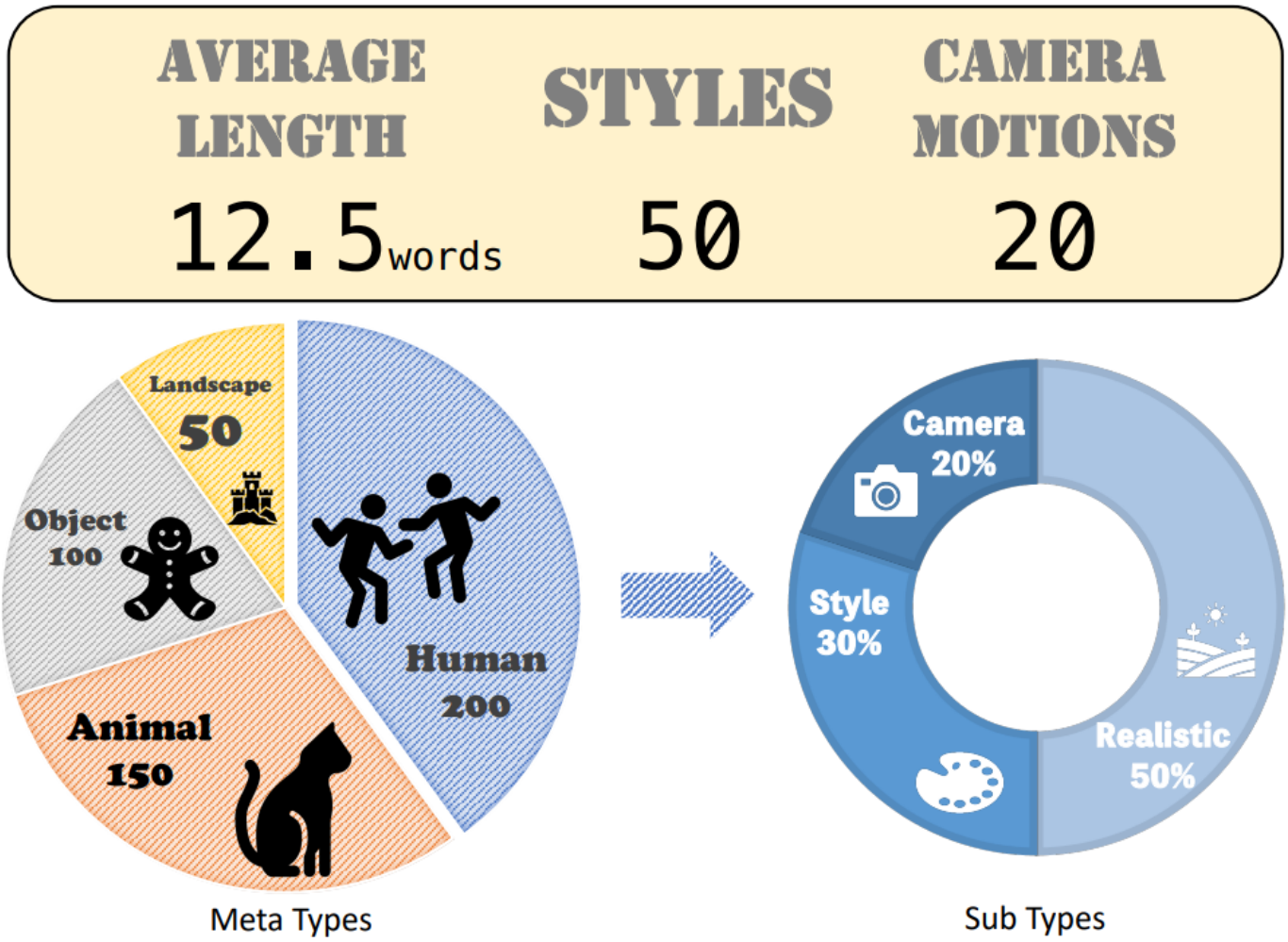

図4：**提案されたベンチマークの分析**。各メタタイプには3つのサブタイプがあり、生成される動画の多様性を高めている。

> The analysis of the proposed benchmarks. Each meta type contains three sub-types to increase the diversity of the generated videos.

全体として、人間、動物、物体、風景のメタクラスで500個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個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。

> Overall, we get over 500 prompts in the meta classes of human, animal, objects, and landscape. Each class contains the natural scenes, the stylized prompts, and the results with **explicit** camera motion controls. We give a brief view of the benchmark in Fig. 4. The whole benchmark contains over 500 prompts with careful categories. To increase the diversity of the prompts, our benchmark contains 3 different sub-types as shown in Figure 4, where we have a total of 50 styles and 20 camera motion prompts. We add them randomly in the 50% prompts of the whole benchmark. Our benchmark contains an average length of 12.5 words pre-prompt, which is also similar to the real-world prompts as we find in Figure. 2.

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	–	–

Method	Ver.	Abilities†	Resolution	FPS	Open Source	Length	Speed	Motion	Camera
VideoCrafter	23.04	T2V	256×256	8	✓	$2s$	0.5 min	–	–
ZeroScope	23.06	T2V&V2V	1024×576	8	✓	$4s$	3 min	–	–
ModelScope-XL	23.08	I2V&V2V	1280×720	8	✓	$4s$	8 min+	–	–
Floor33 Pictures	23.08	T2V	1280×720	8	–	$2s$	4 min	–	–
PikaLab	23.09	I2V or T2V	1088×640	24	–	$3s$	1 min	✓	✓
Gen2	23.09	I2V or T2V	896×512	24	–	$4s$	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。† T2V 手法を主に評価する。関連する I2V モデル、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 I2V を実行する。

> The difference in the available diffusion-based text-to-video models. † We majorly evaluate the method of text-to-video generation (T2V). For related image-to-video generation model (I2V), $i.e.$, ModelScope-XL, we first generate the image by Stable Diffusion v2.1 and then perform image-to-video on the generated content.

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。

> Different from previous FID [46:1] based evaluation metrics, we evaluate the T2V models in different aspects, including the visual quality of the generated video, the text-video alignment, the content correctness, the motion quality, and temporal consistency. Below, we give the detailed metrics.

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 [1:6] 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。

> We first consider the visual quality of the generated video, which is the key for visually appealing to the users. Notice that, since the distribution-based method, $e.g.$, FVD [1:7] still needs the ground truth video for evaluation, we argue these kinds of metrics are not suitable for the general T2V generation cases.

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。

> We utilize the state-of-the-art video quality assessment method, Dover [47:2], to evaluate the quality of the generated video in terms of aesthetics and technicality, where the technical rating measures the quality of the generated video in terms of the common distortion, including noises, artifacts, etc. Dover [47:3] is trained on a self-collected larger-scale dataset and the labels are ranked by the real users for alignment. We term the aesthetic and technical scores as VQA$_A$ and VQA$_T$, respectively.

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 [2:4] 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 [15:2] 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。

> Following previous metrics in the T2V generation papers, we also use the inception score [2:5] of the video as one of the video quality assessment indexes. The inception score is proposed to evaluate the performance of GAN [15:3], which utilizes a pre-trained Inception Network [49:1] on the ImageNet [48:1] dataset as the pre-trained feature extraction method. The inception score reflects the diversity of the generated video, whereas a larger score means the generated content is more diverse.

4.2 テキストと動画の一貫性

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一貫性である。私たちは、グローバルなテキストプロンプトと動画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する。以下に各スコアの詳細を記す。

> Another common evaluation direction is the alignment of the input text and the generated video. We not only consider both the global text prompts and the video, and also the content correctness in different aspects. Below, we give the details of each score.

テキストと動画の整合性（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動画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 [21:4] 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、それらのコサイン類似度を計算する。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 のコサイン類似度は、$emb(\cdot)$ を CLIP 埋め込みとして、$\mathcal{C}\left(emb\left(x_t^i\right), emb\left(p^i\right)\right)$ と表せる。最終的な CLIP-Score（S_{CS}）は、すべてのフレームと動画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 はテスト動画の総数、N は各動画のフレームの総数である。

> We incorporate the CLIP-Score as one of the evaluation metrics, given its widespread usage and simplicity in quantifying the **discrepancy** between input text prompts and generated videos. Utilizing the pretrained ViT-B/32 CLIP model [21:5] as a feature extractor, we obtain frame-wise image embeddings and text embeddings, and compute their cosine similarity. The cosine similarity for the t-th frame of the i-th video x_t^i and the corresponding prompt p^i is denoted as $\mathcal{C}\left(emb\left(x_t^i\right), emb\left(p^i\right)\right)$, $emb(\cdot)$ means CLIP embedding. The overall CLIP-Score, S $_C S$, is **derived** by averaging **individual** scores across all frames and videos, calculated as
>
> $$S_{CS} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N}\sum_{t=1}^{N}\mathcal{C}\left(emb\left(x_t^i\right), emb\left(p^i\right)\right)\right), \tag{1}$$
>
> where M is the total number of testing videos and N is the total number of frames in each video.

テキストと動画の整合性（SD-Score）

現在の動画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。そのため、生成物の品質をフレーム単位で Stable Diffusion [4:4] と比較するという新しい指標を提案する。具体的には、SDXL [5:4] 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、生成された画像と動画の1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、T2I 拡散モデルを T2V 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-Score は以下のようになる。

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

> Most current video diffusion models are fine-turned on a base stable diffusion with a larger scale dataset. Also, tuning the new parameters for stable diffusion will cause conceptual forgetting, we thus propose a new metric by comparing the generated quality with the frame-wise stable diffusion [4:5]. In detail, we use SDXL [5:5] to generate N_1 images $\{dk\}_{k=1}^{N1}$ for every prompt and extract the visual embeddings in both generated images and video frames, and here we set N_1 to 5. We calculate the embedding similarity between the generated videos and the SDXL images, which is helpful to ablate the concept forgotten problems when fine-tuneing the text-to-image diffusion model to video models. The final SD-Score is

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

テキストと動画の整合性（BLIP-BLEW）

また、生成された動画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。T2I の評価方法 [44:2] 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。

> We also consider the evaluation between the text descriptions of the generated video and the input text prompt. To this purpose, we utilize BLIP2 [50:1] for caption generation. Similar to text-to-image evaluation methods [44:3], we use BLEU [51:1] for text alignment of the generated and the source prompt across frames:
>
> $$S_{BB} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N_2}\sum_{k=1}^{N_2}\mathcal{B}\left(p^i, l_k^i\right)\right), \tag{3}$$
>
> where $B(\cdot,\cdot)$ is the BLEU similarity scoring function, $\{l_k^i\}_{k=1}^{N_2}$ are BLIP generated captions for i-th video, and N_2 is set to 5 experimentally.

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、我々が関心を持つ動画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})：動画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、σ_t^i は動画 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})：

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

   として計算されるオブジェクト数の平均の差を評価する。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 は動画 i の検出されたオブジェクトカウントフレーム t、$\hat{c}^i$ は動画 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 は動画 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。

> For general objects, we employ a state-of-the-art segmentation and tracking method, namely SAM-Track [52:1], to analyze the correctness of the video content that we are interested in. Leveraging the powerful segmentation model [53:1], we can easily obtain the objects and their

attributes. In our pipeline, we focus on detecting prompts with COCO classes [54:1], which is a widely used dataset for object detection and segmentation tasks. We evaluate T2V models on the existence of objects, as well as the correctness of color and count of objects in text prompts. Specifically, we assess the Detection-Score, Count-Score, and Color-Score as follows:

1. $Detection\text{-}Score(S_{Det})$: Measures average object presence across videos, calculated as:

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

   where M_1 is the number of prompts with objects, and σ_j^i is the detection result for frame t in video i (1 if an object is detected, 0 otherwise).
2. $Count\text{-}Score(S_{Count})$: Evaluates average object count difference, calculated as:

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

   where M_2 is the number of prompts with object counts, c_t^i is the detected object count frame t in video i and $\hat{c}^i$ is the ground truth object count for video i.
3. $Color\text{-}Score(S_{Color})$: Assesses average color accuracy, calculated as:

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

   where M_3 is the number of prompts with object colors, s_t^i is the color accuracy result for frame i in video t (1 if the detected color matches the ground truth color, 0 otherwise).

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、生成された動画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であり、N_3 は 3 に設定される。

Human is important for the generated videos as shown in our collected real-world prompts. To this end, we also evaluate the correctness of human faces using DeepFace [55:1], a popular face analysis toolbox. We do the analysis by calculating the distance between the generated celebrities' faces with corresponding real images of the celebrities.

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

where M_4 is the number of prompts that contain celebrities, $D(\cdot, \cdot)$ is the Deepface's distance function, $\{f_k^i\}_{k=1}^{N_3}$ are collected celebrities images for prompt i, and N_3 is set to 3.

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。テキスト生成に用いる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 [44:4] またはマルチモデル LLM 評価法 [12:4] 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。

Another hard case for visual generation is to generate the text in the description. To examine the abilities of current models for text generation, we utilize the algorithms from Optical Character Recognition (OCR) models similar to previous text-to-image evaluation [44:5] or

> multi-model LLM evaluation method [12:5]. Specifically, we utilize PaddleOCR to detect the English text generated by each model. Then, we calculate Word Error Rate (WER) [56:1], Normalized Edit Distance (NED) [57:1], Character Error Rate (CER) [58:1], and finally we average these three score to get the OCR-Score.

4.3 動きの質

動画の場合、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つとみなす。ここでは、以下に紹介する2つの異なる動きの質の評価方法を考える。

> For video, we believe the motion quality is a major difference from other domains, such as image. To this end, we consider the quality of motion as one of the main evaluation metrics in our evaluation system. Here, we consider two different motion qualities introduced below.

行動認識（Action-Score）

人間に関する動画の場合、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 toolbox [59]、特に事前に訓練された VideoMAE V2 [60] モデルを使用して、生成された動画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 400 action classes [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などを含む。

> For videos about humans, we can easily recognize the common actions via pretrained models. In our experiments, we use MMAction2 toolbox [59:1], specifically the pre-trained VideoMAE V2 [60:1] model, to infer the human actions in the generated videos. We then take the classification accuracy (ground truth are actions in the input prompts) as our Action-Score. In this work, we focus on Kinetics 400 action classes [61:1], which is widely used and **encompasses** human-object interactions like playing instruments and human-human interactions, including handshakes and hugs.

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、特定の生成動画クリップの平均 Flow-Score 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ためである。

> We also consider the general motion information of the video. To this end, we use the pretrained optical flow estimation method, RAFT [62:1], to extract the dense flows of the video in every two frames. Then, we calculate the average flow on these frames to obtain the average flow score of every specific generated video clip since some methods are likely to generate still videos which are hard to identify by the temporal consistency metrics.

振れ幅分類スコア（Motion AC-Score）

平均フローに基づき、生成された動画の動きの大きさが、テキストプロンプトで指定された大きさ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 ρ を設定し、ρ を超えると1つの動画が大きいと見なす。ここでは主観的な観察に基づいて ρ を 2 に設定している。生成された動画の動きを識別するために、このスコアをマークする。

> Based on the average flow, we further identify whether the motion amplitude in the generated video is consistent with the amplitude specified by the text prompt. To this end, we set an average flow threshold ρ that if surpasses ρ, one video will be considered large, and here ρ is set to 2 based on our subjective observation. We mark this score to identify the movement of the generated video.

4.4 時間的整合性

時間的整合性もまた、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。

> Temporal consistency is also a very valuable field in our generated video. To this end, we involve several metrics for calculation. We list them below.

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 [62:2] 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。

> We first consider the warping error, which is widely used in previous blind temporal consistency methods [63:1] [64:1] [65:1]. In detail, we first obtain the optical flow of each two frames using the pre-trained optical flow estimation network [62:3], then, we calculate the pixel-wise differences between the warped image and the predicted image. We calculate the warp differences on every two frames and calculate the final score using the average of all the pairs.

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、以前の動画編集作品 [10:8] [65:2] 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。

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

> Besides pixel-wise error, we also consider the semantic consistency between every two frames, which is also used in previous video editing works [10:9] [65:3]. Specifically, we consider the semantic embeddings on each of the two frames of the generated videos and then get the averages on each two frames, which is shown as follows:
>
> $$S_{CT} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N-1}\sum_{t=1}^{N-1}\mathcal{C}\left(emb\left(x_t^i\right), emb\left(x_{t+1}^i\right)\right)\right), \tag{8}$$

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コサイン類似度を計算する。そして、その類似度を平均して最終スコアとする。

$$S_{FC} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N-1}\sum_{t=1}^{N-1}\mathcal{C}\left(emb\left(x_{t+1}^i\right), emb\left(x_1^i\right)\right)\right) \tag{9}$$

> Similar to CLIP-Temp, we evaluate the human identity consistency of the generated videos. Specifically, we select the first frame as the reference and calculate the cosine similarity of the reference frame embedding with other frames' embeddings. Then, we average the similarities as the final score:
>
> $$S_{FC} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N-1}\sum_{t=1}^{N-1}\mathcal{C}\left(emb\left(x_{t+1}^i\right), emb\left(x_1^i\right)\right)\right), \tag{9}$$

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**：生成された動画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. **テキストと動画の位置合わせ**：生成された動画と入力テキストプロンプトの間の関係を考慮し、生成された動画に間違ったカウント、属性、関係がある場合、低品質のサンプルと見なされる。
3. **動きの質**：この指標では、ユーザーは動画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**：時間的整合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。

> Besides the above objective metrics, we conduct user studies on the main five aspects to get the users' opinions. These aspects include (1) $Video\ Qualities$. It indicates the quality of the generated video where a higher score shows there is no blur, noise, or other visual degradation. (2) $Text\ and\ Video\ Alignment$. This opinion considers the relationships between the generated video and the input text-

prompt, where a generated video has the wrong count, attribute, and relationship will be considered as low-quality samples. (3) $Motion\ Quality$. In this metric, the users need to identify the correctness of the generated motions from the video. (4) $Temporal\ Consistency$. Temporal consistency is different from motion quality. In motion quality, the user needs to give a rank for high-quality movement. However, in temporal consistency, they only need to consider the frame-wise consistency of each video. (5) $Subjective\ likeness$. This metric is similar to the aesthetic index, a higher value indicates the generated video generally achieves human preference, and we leave this metric used directly.

Dementions	Metrics	ModelScope-XL [29:2]	ZeroScope [28:2]	Floor33 [8:8]	PikaLab [11:8]	Gen2 [10:10]
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。

Raw results from the aspects of video quality, text-video alignment, motion quality, and temporal consistency.

評価のために、ModelScope [29:3]、ZeroScope [28:3]、Gen2 [10:11]、Floor33 [66]、PikaLab [11:9] 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動画を生成し、合計 2,500 本の動画を得た。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 [5:6] 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、モデルのテキストと映像の整合性を評価するための 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 の間で意見を求める。値が大きいほど良い指標であることを示す。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。

For evaluation, we generate videos using the provided prompts benchmark on five state-of-the-art methods of ModelScope [29:4], ZeroScope [28:4], Gen2 [10:12], Floor33 [66:1], and PikaLab [11:10], getting 2.5k videos in total. For a fair comparison, we change the aspect ratio of Gen2 and PikaLab to 16 : 9 to suitable other methods. Also, since PikaLab can not generate the content without the visual watermark, we add the watermark of PikaLab to all other methods for a fair comparison. We also consider that some users might not understand the prompt well,

> for this purpose, we use SDXL [5:7] to generate three reference images of each prompt to help the users understand better, which also inspires us to design an SD-Score to evaluate the models' text-video alignments. For each metric, we ask three users to give opinions between 1 to 5, where a large value indicates better alignments. We use the average score as the final labeling and normalize it to range $[0, 1]$.

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、上記の個々の指標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）。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。

> Upon collecting user data, we proceed to perform human alignment for our evaluation metrics, with the goal of establishing a more reliable and robust assessment of T2V algorithms. Initially, we conduct alignment on the data using the mentioned individual metrics above to approximate human scores for the user's opinion in the specific aspects. We employ a linear regression model to fit the parameters in each dimension, inspired by the works of the evaluation of natural language processing [67:1] [68:1]. Specifically, we randomly choice 300 samples from four different methods as the fittings samples and left the rest 200 samples to verify the effectiveness of the proposed method (as in Table. 4). The coefficient parameters are obtained by minimizing the residual sum of squares between the human labels and the prediction from the linear regression model. In the subsequent stage, we integrate the aligned results of these four aspects and calculate the average score to obtain a comprehensive final score, which effectively represents the performance of the T2V algorithms. This approach streamlines the evaluation process and provides a clear indication of model performance.

Aspects	**Methods**	**Spearman'sρ**	**Kendall'sϕ**
Visual Quality	VQA_A	42.1	30.5
	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**。T2V 変換におけるいくつかの客観的指標と人間の評価との相関。相関計算にはスピアマンの ρ とケンドールの ϕ を使用する。

> Correction Analysis. Correlations between some objective metrics and human judgment on text-to-video generations. We use Spearsman's ρ and Kendall's ϕ for correlation calculation.

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 [29:5]、Floor33 Pictures [66:2]、ZeroScope [28:5] 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 [10:13] と PikaLab [11:11] 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動画を生成する。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。

> We conduct the evaluation on 500 prompts from our benchmark prompts, where each prompt has a metafile for additional information as the answer of evaluation. We generate the videos using all available high-resolution T2V models, including the ModelScope [29:6], Floor33 Pictures [66:3], and ZeroScope [28:6]. We keep all the hyper-parameters, such as classifier-free guidance, as the default value. For the servicebased model, we evaluate the performance of the representative works of Gen2 [10:14] and PikaLab [11:12]. They generate at least 512p videos with high-quality watermark-free videos. Before our evaluation, we show the differences between each video type in Table 1, including the abilities of these models, the generated resolutions, and fps. As for the comparison on speed, we run all the available models on an NVIDIA A100. For the unavailable model, we run their model online and measure the approximate time. Notice that, PikaLab [11:13] and Gen2 [10:15] also have the ability to control the motions and the cameras through additional hyper-parameters. Besides, although there are many parameters that can be adjusted, we keep the default settings for a relatively fair comparison.

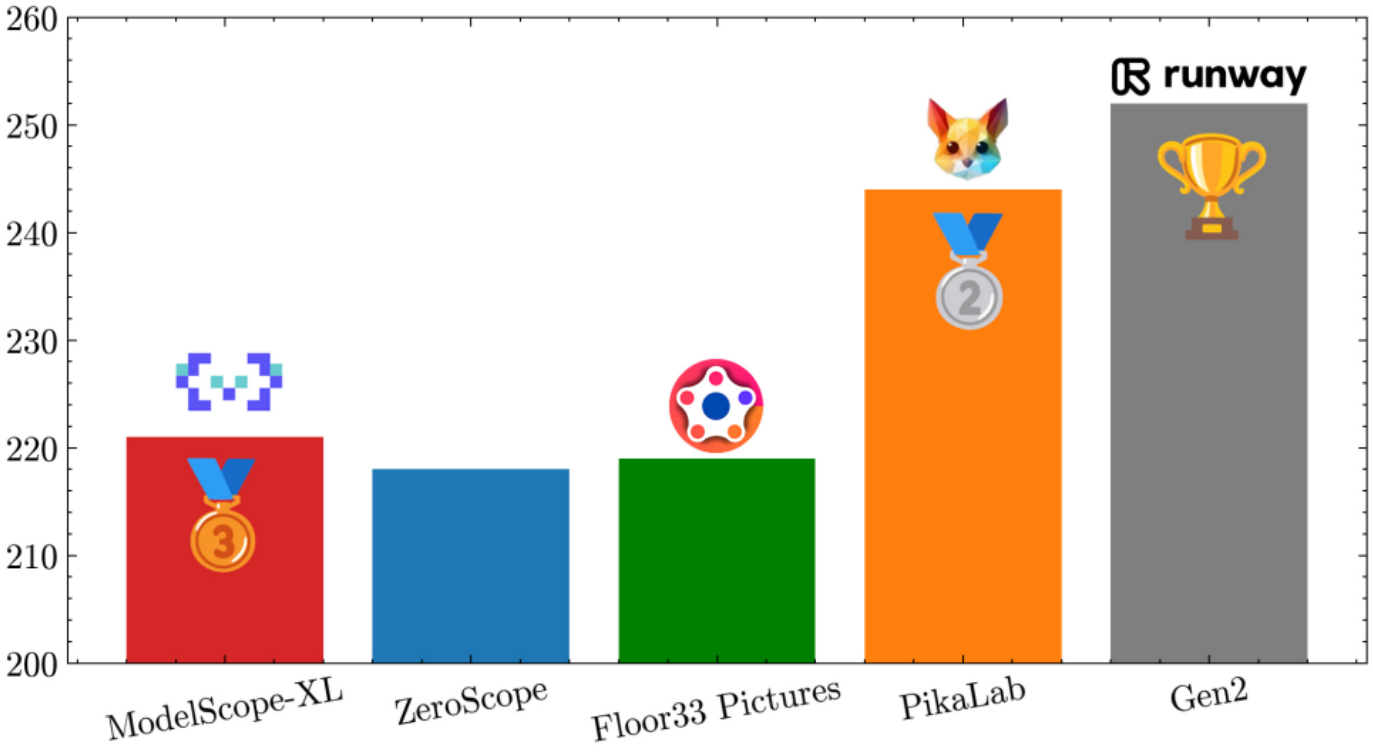

図5：EvalCrafterベンチマークでの総合比較結果。

> Overall comparison results on our EvalCrafter benchmark.

	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)

	Visual Quality	Text-Video Alignment	Motion Quality	Temporal Consistency
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）。

> Human-preference aligned results from four different aspects, with the rank of each aspect in the brackets.

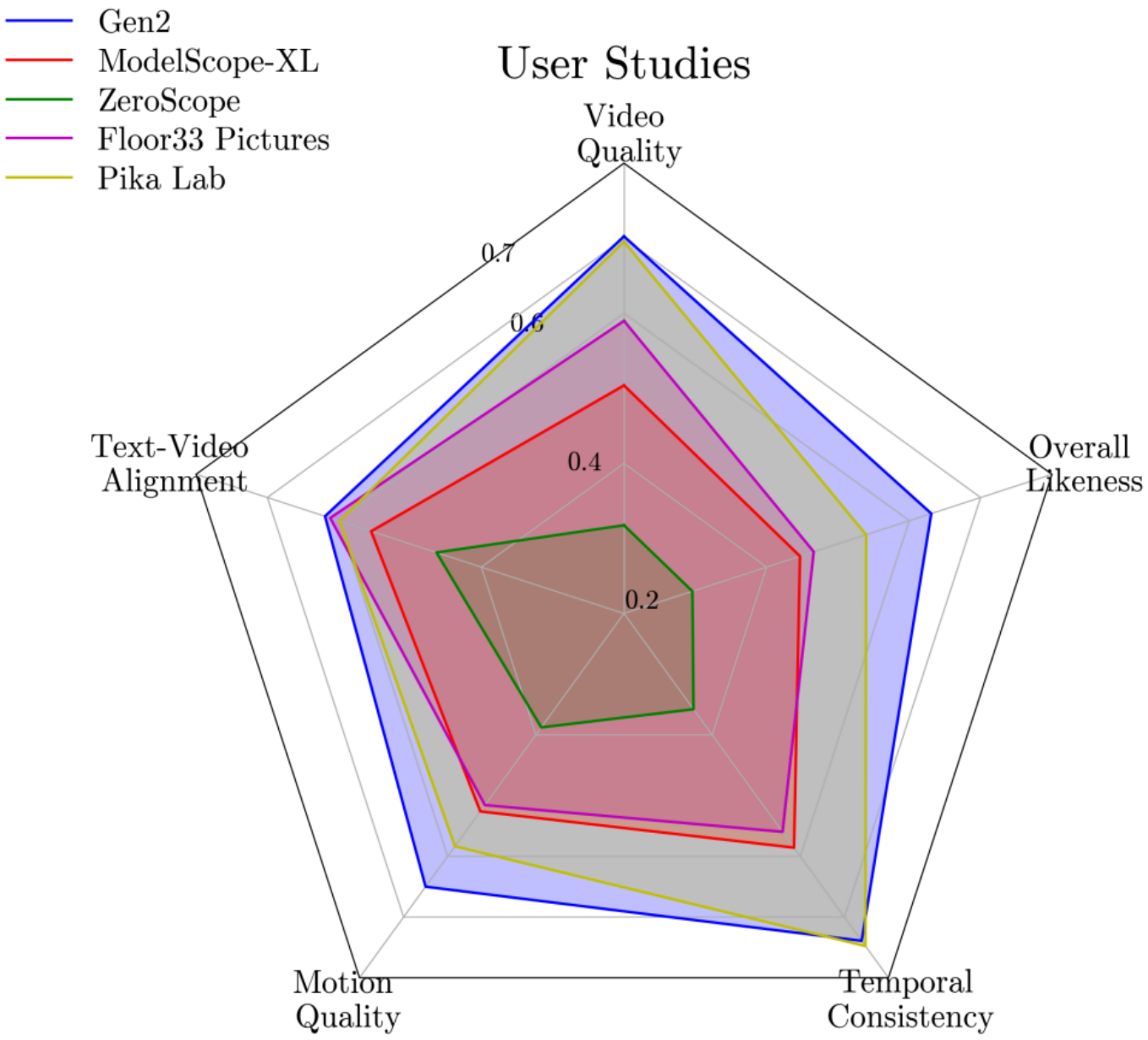

図6：**ユーザー調査による生の評価。**

> The raw ratings from our user studies.

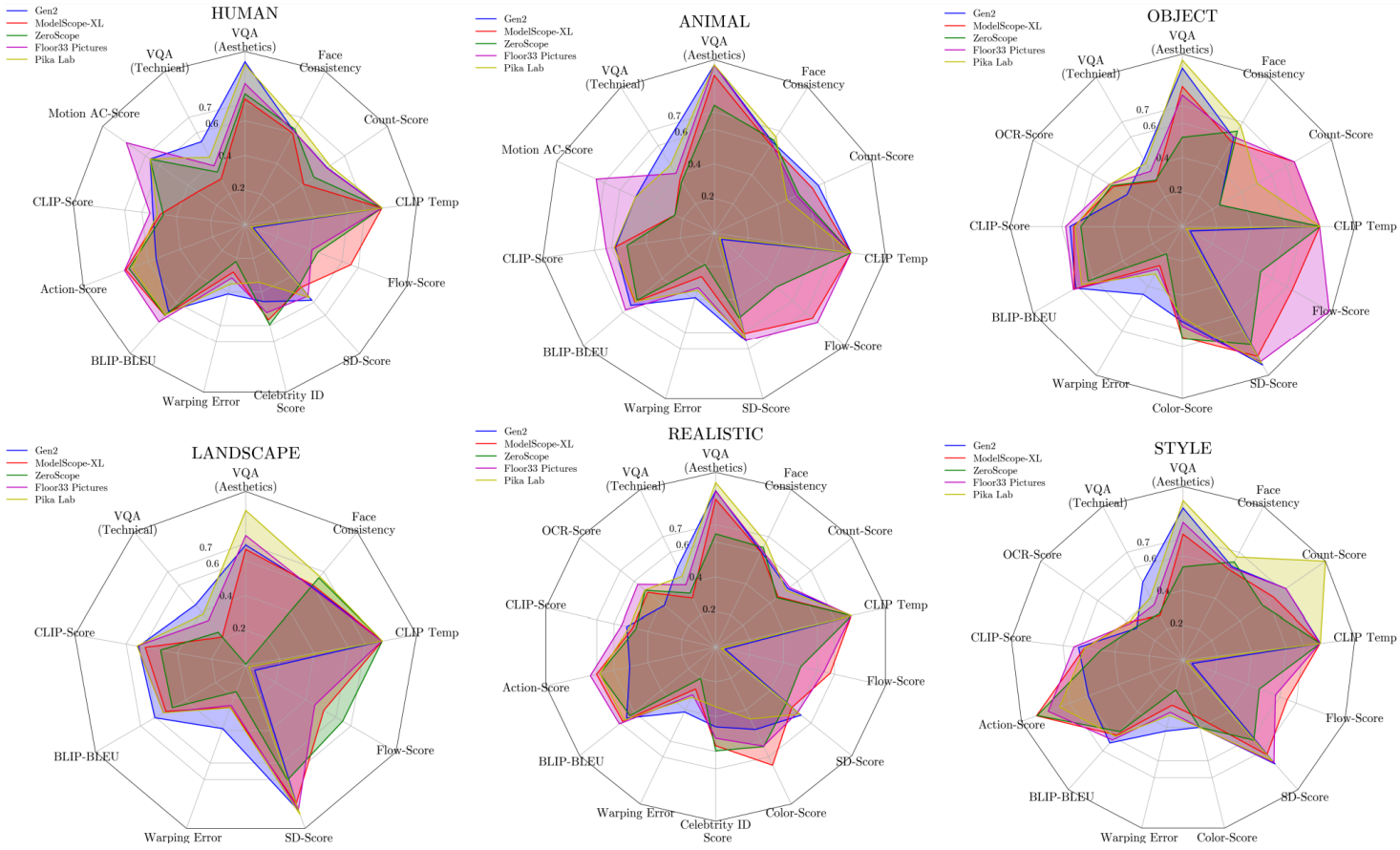

図7：さまざまな側面における生の結果。生成された動画のメタタイプの性能を評価するために、4つの主要なメタタイプ（動物、人間、風景、物体）を考慮する。各タイプには、きめ細かい属性ラベルを持つ複数のプロンプトが含まれる。それぞれのプロンプトに対して、動画のスタイルも考慮するが、さらに、上の現実的でスタイリッシュな図のように、多様なプロンプトがある。(メトリクスの値は、より見やすくするために正規化されている。ワープエラーとOCRスコアを前処理しているため、この図ではこの2つのメトリクスの値が大きいほど性能が良いことを示している)

> *Raw results in different aspects*. We consider 4 main meta types (animal, human, landscape, object) to evaluate the performance of the meta types of the generated video, where each type contains several prompts with fine-grained attribute labels. For each prompt, we also consider the style of the video, yet more diverse prompts. as shown in realistic and style figure above. (The metrics values are normalized for better visualization, we preprocess the warping Error and OCR-score, so for these two metrics, a large value indicates better performance in this figure.)

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。

> We first show the overall human-aligned results in Fig. 5, with also the different aspects of our benchmark in Table 3, which gives us the final and the main metrics of our benchmark. Finally, as in Figure 7, we give the results of each method on four different meta-types (*i.e.*, animal, human, landscape, object) in our benchmark and two different type videos (*i.e.*, general, style) in our benchmark. For comparing the objective and subjective metrics of each method, we give the raw data of each metric in Table. 1 and Fig. 6. We give a detailed analysis in Sec. 5.1.

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わかった。例えば、Gen2 が映像の質、動きの質、時間的一貫性の点で他のモデルを凌駕する一方で、PikaLab はテキストと動画の位置合わせで優れたパフォーマンスを示している。

> From Table. 3, the rankings of the models **vary** significantly across these aspects, highlighting the importance of a multi-aspect evaluation approach for a comprehensive understanding of their performance. For instance, while Gen2 outperforms other models in terms of Visual Quality, Motion Quality, and Temporal Consistency, PikaLab demonstrates superior performance in Text-Video Alignment.

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 [10:16] 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。主観的には、Gen2 [10:17] 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、より弱い動きの振れ幅で生成していると考察する。

> As shown in Fig. 7, most methods show very different values in different meta types. For example, although Gen2 [10:18] has the best overall T2V alignment in our experiments, the generated videos from this method are hard to recognize by the action recognition models. We subjectively find Gen2 [10:19] mainly generates the close-up shot from text prompt with a weaker motion amplitude.

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 [10:20] はすべてのテキストと映像の整合性測定基準において良い結果を出すことはできないが、その良好な時間的整合性、視覚的品質、および小さな振れ幅により、ユーザーはほとんどのケースでこのモデルの結果を受け入れる。

> As shown in Fig. 7 and Table. 2, even Gen2 [16] can not perform well in all the text-video alignment metrics, the user still likes the results of this model in most cases due to its good temporal consistency, visual quality, and small motion amplitude.

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。

> Although some additional hyper-parameters can be set as additional control handles, the text encoder of the current T2V text encoder still lacks the understanding of the reasoning behind open-world prompts, like camera motion.

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 [10:21] 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。

> As shown in Tab. 1, Gen2 [10:22] has the smallest resolutions, however, both humans and the objective metrics consider this method to have the best visual qualities and few artifacts as in Tab. 2, Fig. 6.

発見⑥ 動きの振れ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、2つの小さなモーションモデルである PikaLab [11:14] と Gen2 [10:23] は、大きなモーションモデルである Floor33 Pictures [66:4] 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。

> From Fig. 6, both two small motion models, $i.e.$, PikaLab [11:15] and Gen2 [10:24] get better scores in the user's choice than the larger motion model, $i.e.$, Floor33 Pictures [66:5]. Where users are more likely to see slight movement videos other than a video with bad and unreasonable motions.

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。

> Although we report the OCR-Scores of these models, we find it is still too hard to generate realistic fonts from the text prompts, nearly all the methods are fair to generate high-quality and consistent texts from text prompts.

発見⑧ 現在の動画生成モデルは、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値を示している。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。

> All methods show a very high consistency of CLIP-Temp as in Table. 2, which means each frame has a very similar semantic across frames. So the current T2V models are more likely to generate the cinemagraphs, other than the long video with multiple transitions and actions.

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映像の整合性に価値がある。VQA_T と VQA_A は、視覚的な品質評価にも価値がある。

> By aligning the objective metrics to the real users, we also find some valuable metrics from a single aspect. For example, SD-Score and CLIP-Score are both valuable for text-video alignment according to Table. 2 and Table. 3. VQA_T and VQA_A are also valuable for visual quality assessment.

発見⑩ Gen2 も完璧ではない

Gen2 [10:25] 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動画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。

> Although Gen2 [10:26] achieved the overall top performance in our evaluation, it still has multiple problems. For example, Gen2 [10:27] is hard to generate video with complex scenes from prompts. Gen2 [10:28] has a weird identity for both humans and animals, which is also reflected by the IS metric (hard to be identified by the network also) in Table. 1, while other methods do not have such problems.

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 [11:16] や Gen2 [10:29] 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。

> Referring to Table 3, we can observe that open-source models such as ModelScope-XL and ZeroScope have lower scores in almost every aspect compared to closed-source models like PikaLab [11:17] and Gen2 [10:30]. This indicates that there is still room for improvement in open-source T2V models to reach the performance levels of their closed-source counterparts.

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、現在の動きの振れ幅のスコアはすべて、ユーザーの選択とは関係がないということだ。私たちは、人間は振れ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。

> To demonstrate the effectiveness of our model in aligning with human scores, we calculate Spearman's rank correlation coefficient [69:1] and Kendall's rank correlation coefficient [70:1], both of which are non-parametric measures of rank correlation. These coefficients provide insights into the strength and direction of the association between our method results and human scores, as listed in Table. 4. From this table, the proposed weighting method shows a better correlation on the unseen 200 samples than directly averaging (we divide all data by 100 to get them to range $[0, 1]$ first). Another interesting finding is that all current Motion Amplitude scores are not related to the users' choice. We argue that humans care more about the stability of the motion than the amplitude. However, our fitting method shows a higher correlation.

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個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 や大規模な動画基礎モデルの時代には、より優れたより大規模な動画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。

> Although we have already made a step in evaluating the T2V generation, there are still many challenges. (i) Currently, we only collect 500 prompts as the benchmark, where the real-world situation is very complicated. More prompts will show a more detailed benchmark. (ii) Evaluating the motion quality of the general senses is also hard. However, in the era of multi-model LLM and large video foundational models, we believe better and larger video understanding models will be released and we can use them as our metrics. (iii) The labels used for alignment are collected from only 3 human annotators, which may introduce some bias in the results. To address this limitation, we plan to expand the pool of annotators and collect more diverse scores to ensure a more accurate and unbiased evaluation.

6 結論

オープンワールドの大規模な生成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、テキストと映像の整合性、対象物、動きの質について、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ー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。

> Discovering more abilities of the open world large generative models is essential for better model design and usage. In this paper, we make the very first step for the evaluation of the large and high-quality T2V models. To achieve this goal, we first built a detailed prompt benchmark for T2V evaluation. On the other hand, we give several objective evaluation metrics to evaluate the performance of the T2V models in terms of the video quality, the text-video alignment, the object, and the motion quality. Finally, we conduct the user study and propose a new alignment method to match the user score and the objective metrics, where we can get final scores for our evaluation. The experiments show the abilities of the proposed methods can successfully align the users' opinions, giving the accurate evaluation metrics for the T2V methods.

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. Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022. ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩
9. 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. ↩ ↩ ↩ ↩ ↩ ↩
10. 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. ↩
11. Pika Lab discord server. https://www.pika.art/. Accessed: 2023-08-30. ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩ ↩
12. Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. Jul 2023. ↩ ↩ ↩ ↩ ↩ ↩
13. 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. ↩ ↩ ↩ ↩
14. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013. ↩ ↩
15. 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. ↩ ↩ ↩ ↩
16. 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. ↩ ↩
17. 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. ↩ ↩
18. 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. ↩ ↩
19. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020. ↩ ↩
20. Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 2 ↩ ↩
21. 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. ↩ ↩ ↩ ↩ ↩ ↩
22. 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. ↩ ↩ ↩ ↩ ↩ ↩
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. ↩ ↩