T2VBench: Benchmarking Temporal Dynamics for Text-to-Video Generation

Pengliang Ji[1], Chuyang Xiao[2]*, Huilin Tai[3], Mingxiao Huo[1]
[1]Carnegie Mellon University [2]ShanghaiTech University [3]McGill University
{pengliaj, mingxiah}@andrew.cmu.edu, xiaochy@shanghaitech.edu.cn, huilin.tai@mail.mcgill.ca

Abstract

While text-to-video (T2V) generative models produce exceptionally realistic videos, they lack a comprehensive evaluation across the temporal dimension, with a limited focus on basic dynamics including camera transitions, movement, and event sequences. In this work, we introduce **T2VBench**, *a comprehensive T2V evaluation benchmark enriched with temporal dynamics lexicons derived from curated temporal words on Wikipedia. T2VBench is a hierarchical evaluation framework comprising over 1,600 temporally rich prompts and 5,000 generated videos with human ratings, spanning 16 critical temporal evaluation dimensions. We assess three leading text-to-video models, including ZeroScope and Pika, to gauge their proficiency in handling temporal dynamics. Our analysis highlights the strengths and limitations of these models across various temporal aspects. Furthermore, we provide insights into future directions for enhancing text-to-video evaluation metrics and offer a detailed analysis of these models' performance across the temporal dimensions. Overall, T2VBench is the* **first-of-its-kind** *comprehensive benchmark fully focused on temporal dynamics for text-to-video evaluation. It aims to facilitate scientific benchmarking of both generative models and automated metrics on text-to-video generation.*

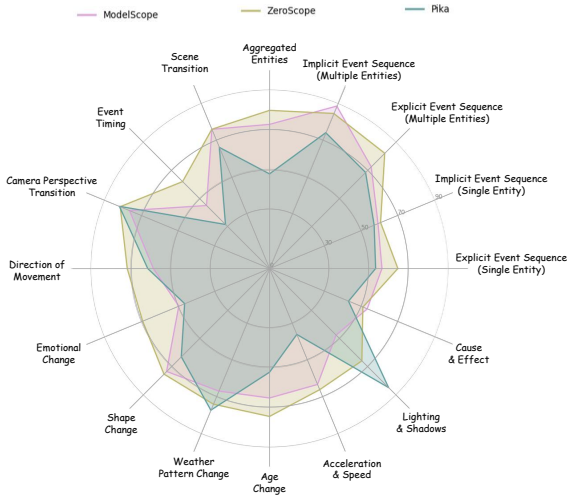

Figure 1. **Leaderboard of Text-to-Video Models on T2VBench.** Our hierarchical evaluation benchmark offers an in-depth analysis of model performance across the temporal dimension, encompassing a wide range of dynamics. This comprehensive assessment includes a detailed examination of 16 specific aspects for text-to-video (T2V) evaluation. To facilitate comparison, all scores are normalized to a maximum of 100.

1. Introduction

Generative models have seen significant advancements in recent years across various domains, such as computer vision [1, 2], robotics [3, 4], and scientific fields [5]. In the realm of computer vision, diffusion models have propelled the field of text-to-video (T2V) generation forward, enabling the creation of customized videos directly from textual prompts. This has been demonstrated by pioneering works such as those mentioned in [6–10]. OpenAI's Sora [11], which utilizes diffusion transformers [12], represents a landmark development with its unparalleled ability to generate videos. These videos are not only realistic but also capable of robust physical simulations. The emergence of T2V generative models necessitates the development of comprehensive evaluation benchmarks to assess the quality of these models more thoroughly.

Expanding the Horizon of Text-to-Video Evaluation. Prior works have laid the foundation for benchmarks assessing the quality of text-to-video synthesis. EvalCrafter [13] took an initial step by establishing a pipeline for generation and incorporating human ratings to regularize the evaluation of these models. Subsequently, FETV [14] advanced the field significantly by introducing attribute control in both spatial and temporal dimensions, resulting in a more comprehensive and meaningful benchmark. However, the release of Sora has revealed that current text-to-video models still struggle with temporal consistency, such as the unexpected disappearance of objects. Therefore, it is crucial to

*Equal Contribution, Project Page: https://ji-pengliang.github.io/T2VBench/

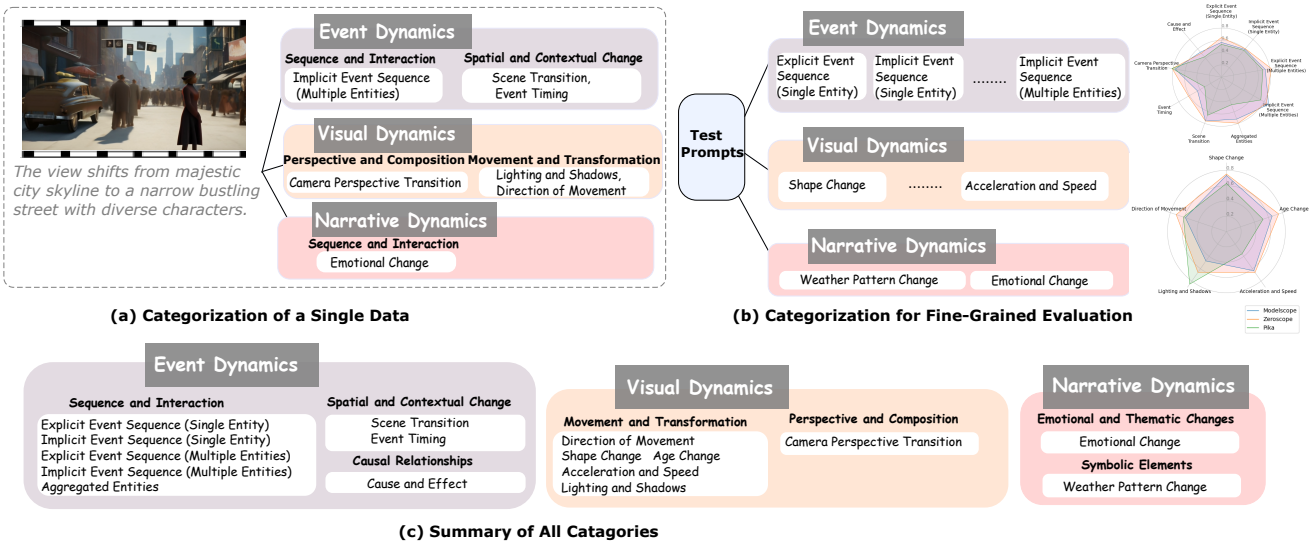

Figure 2. **Illustration of the Fine-grained Structure for Crafting Evaluation Dimensions.** The structure of T2VBench organizes lexicons across three tiers: Event Dynamics, Visual Dynamics, and Narrative Dynamics. Each tier is intricately subdivided into 7 and 16 aspects, respectively, offering an exhaustive framework for the evaluation criteria applicable to a wide array of samples.

introduce benchmarks that focus more on capturing temporal consistency. In this paper, we provide a more refined prompting approach for the temporal dimension and introduce a benchmark that more effectively represents temporal coherence in text-to-video evaluation.

A comprehensive benchmark across temporal dynamics. Current benchmarks primarily focus on spatial aspects and basic temporal tasks such as the sequencing of events, failing to provide a detailed and comprehensive evaluation of Text-to-Video (T2V) models. To address this issue, we have compiled a *word bag* from Wikipedia and developed a temporal lexicon rich in temporal dynamics. Furthermore, we have introduced an automated process to generate a wide variety of prompts, each reflecting different types of temporal change. This process allows for the creation of prompts that encompass a broad spectrum of temporal dynamics, facilitating a more nuanced evaluation of Text-to-Video models. As a result, we have collected 1,680 prompts that are rich in temporal information, enabling an extensive evaluation. These prompts are supplemented with thorough annotations based on human preferences, further enhancing the value of our benchmark.

Assess models and automatic metrics on T2VBench. Upon prompting for temporal text and video generation, we evaluate our models using a suite of rich text-to-video evaluation metrics beyond human ratings. Our benchmarks are appraised using classical text-to-vision scores such as the CLIPScore [15] and the BLIPScore [16], in addition to metrics that utilize human feedback, including ImageReward [17], PickScore [18], and HPSv2 [19]. To our knowledge, we are the first to assess our benchmark using the VQAScore [20]. The VQAScore represents a significant advancement, leveraging Vision-Language Models (VLMs) such as LLaVA [21] and InstructBLIP [22], which can directly and comprehensively understand visual content. However, as text-to-video models advance and prompts become more flexible, there is a growing need for modern evaluation metrics that assess the consistency between the generated video and lengthy text prompts. This drives the development of new metrics aligned with the evolving requirements.

In summary, our contributions are as follows:

- We introduce **T2VBench**, the first hierarchical and comprehensive benchmark specifically designed for temporal dynamics in text-to-video evaluation, encompassing over 1,600 prompts and 5,000 generated videos enriched with lexicons that effectively capture temporal dynamics.
- We conduct a thorough analysis of the most popular T2V models, examining their capabilities across various evaluation aspects and scrutinizing their performance in handling 16 critical aspects of temporal dynamics.
- We evaluate automatic metrics for their effectiveness in temporal assessment within our benchmark and offer insightful recommendations for the evaluation of generative foundation models (GenFMs).

2. Related Works

2.1. Evaluation on Text-to-video Generative Models

Alignment and quality are two main streams of text-to-visual evaluation. While human feedback excels in preci-

Table 1. **Comparing Temporal Evaluation Skill Coverage Across Benchmarks.** In comparison to existing text-to-video benchmarks, our benchmark provides fine-grained coverage of all essential temporal skills across a range of temporal dynamics, particularly in advanced and complex scenarios, laying the foundation for a comprehensive text-to-video evaluation.

Benchmarks	Event Dynamics					Visual Dynamics						Narrative Dynamics	
	Event Sequence	Aggregated Entities	Scene Transition	Event Timing	Cause & Effect	Camera Perspective Transition	Direction of Movement	Geometry Change	Age Change	Acceleration &Speed	Lighting & Shadows	Emotional Change	Weather Pattern Change
EvalCrafter [13]	✓	✗	✓	✗	✗	✓	✗	✗	✗	✓	✓	✓	✗
T2VScore [23]	✓	✗	✓	✗	✗	✓	✗	✗	✗	✓	✓	✓	✗
FETV [14]	✓	✗	✗	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗
Video-Bench [24]	✓	✓	✓	✗	✗	✓	✓	✓	✗	✓	✓	✓	✗
VBench [25]	✓	✗	✗	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗
MSR-VTT [26]	✓	✗	✗	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

sion, it comes with considerable expense. In contrast, automated metrics [27–29], such as the Video Inception Score and Kernel Video Distance, have been introduced for video quality evaluation. However, evaluation of Text-to-Video alignment currently has only a few methods [23], which remain under construction. Encouragingly, the advent of vision-language foundation models (VLMs) has brought image-based evaluation metrics [20, 30–40], such as CLIPScore, BLIPScore, and VQAScore, into focus. These metrics enhance the assessment of generative videos through their superior ability to comprehensively gauge alignment.

2.2. Benchmarks for Text-to-Video Generation

The advancement of text-to-video generative models [8, 11, 41–43] has spurred the development of specialized evaluation benchmarks. Recently proposed benchmarks, such as EvalCrafter [13], represent pioneering efforts in assessing large T2V models, focusing on video quality, consistency, and alignment with the input text. Following this, FETV [44] introduced a hierarchical approach to crafting prompts that encapsulate fundamental aspects of spatial and temporal evaluation, such as motion and visual storytelling. T2VScore [23] further refines this approach by leveraging EvalCrafter's prompts to enhance annotations with a particular emphasis on video quality. Despite these advancements, current benchmarks fail to evaluate the comprehensive temporal dynamics in generative videos, a gap highlighted by the analysis in Tab.1. To our knowledge, T2VBench, rooted in a carefully crafted temporal lexicon, stands out as a pioneering benchmark aimed at comprehensively assessing the temporal dimension in text-to-video generative models.

3. T2VBench: A Comprehensive Hierarchical Benchmark for Text-to-video Evaluation

In this section, we introduce the insights behind the construction of our benchmark and pipeline, as shown in Fig. 2, and provide statistics to highlight its effectiveness for comprehensive text-to-video evaluation.

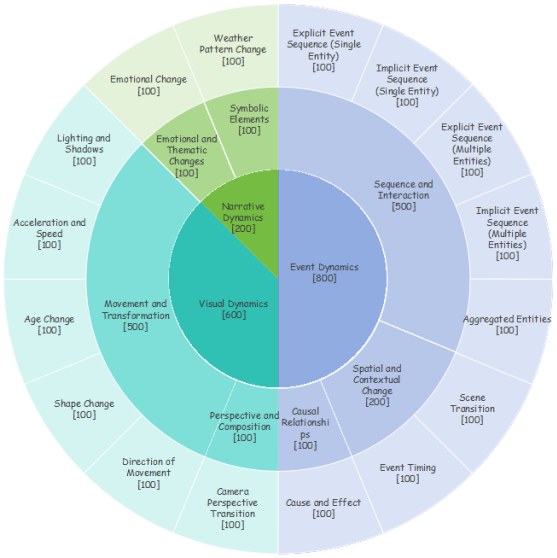

Figure 3. **Hierarchical Structure of Evaluation Aspects**. Leveraging a meticulously designed hierarchical lexicon for assessing text-to-video generation, we've amassed a collection of hierarchical evaluation perspectives spanning the temporal dimension. Our benchmarks encompass a majority of event dynamics, followed by visual dynamics, mirroring the distribution found in the real world. Each category is further divided into fine-grained classes, ensuring an average of 100 samples corresponding to each aspect.

3.1. Towards Temporal Dynamics

Lexicons rich in temporal dynamics. Comprehensive evaluation is driven by fine-grained evaluation dimensions. Here, we propose a keyword-based depth-first search recursion method on Wikipedia to identify time-related effects in the form of words, phrases, and sentences. Over 5,000 entries are collected through this process. We then curate these collected data by manually removing duplicates based on semantics and their importance across real-world distribution. Finally, we compile an evaluation word bag covering 323 critical temporally-related words and phrases for text-to-video evaluation. This word bag consists of temporal ele-

ments including consistency, logic, interaction, simulation, etc. The evaluation word bag will be published alongside our benchmark to facilitate further research and analysis in the field.

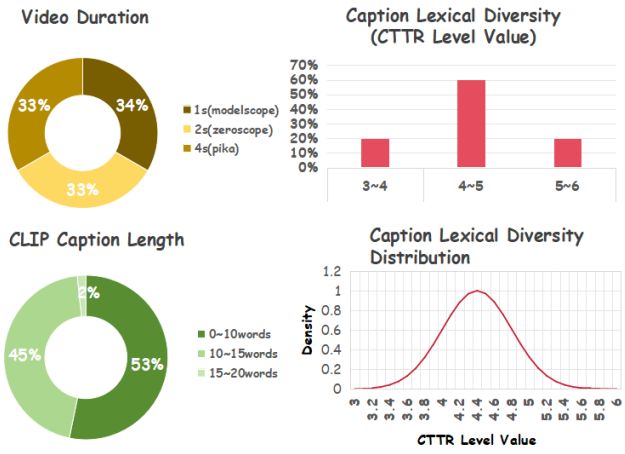

Figure 4. **T2VBench is a Temporally Diverse, Semantically-Rich Benchmark for Text-to-Video Evaluation.** Analyzing captions and video lengths reveals our benchmark's significant semantic diversity, shown by a high CTTR, covering captions from 5 to 30 words to reflect real-world video descriptions. This ensures a broad temporal exploration for comprehensive T2V evaluation.

Evaluation Dimension Crafting. Leveraging lexicons rich in temporal dynamics and semantics [45–53], we present a hierarchical structure through fine-grained classification. The lexicon is segmented into three levels, considering the frequency and temporal characteristics of real-world distributions, as illustrated in Fig.3. These levels encompass *Event Dynamics*, *Visual Dynamics*, and *Narrative Dynamics*. Following the principle of comprehensively covering evaluation aspects, the second and third levels are further divided using the same methodology into 7 and 16 aspects, respectively, as shown in Tab. 2.

3.2. Benchmark Construction

Temporal Prompts Creation. Grounded in the meticulously crafted evaluation dimensions, we curate prompts rich in temporal semantics to thoroughly assess text-to-video generation models. We manually create 100 prompts for each primary evaluation aspect, ensuring that each prompt encompasses 2 to 5 sub-evaluation aspects from at least two secondary categories. Subsequently, we evaluate the prompts' alignment with real-world scenarios to refine our selection. To avoid prompts that are open to diverse interpretations, which could lead to inaccurate evaluations, we carefully guide the designers in prompt creation. Ultimately, we compile 1,680 semantically rich prompts, encapsulating diverse temporal dynamics for a comprehensive evaluation of text-to-video capabilities.

Visual Content Generation. We evaluate three of the most popular text-to-video models: ModelScope [8], Zeroscope [42], and Pika [41]. Utilizing the collected prompts, we generate videos through their respective APIs or websites. To ensure the stability and reliability of the generated content, we create five videos for each prompt and calculate their average performance across various metrics.

Human Ratings Collection. To assist and validate the reliability of automatic metrics within our text-to-video benchmark, we gather human preferences for the generated videos. We employ a 1-5 Likert scale to collect human ratings, adhering to a specified annotation protocol [54]. Each (prompt, video) pair is evaluated by three annotators, with their average rating serving as the final score. Throughout this process, we monitor the variance among annotator scores on identical samples to ensure the quality of the annotations. A concise overview of our benchmark is depicted in Fig. 4. The benchmark features an average prompt length of 15.7 words, with video durations averaging three seconds, aligning with the existing upper limit of high-quality, open-source T2V model capabilities. Additionally, over 80% of our prompts exhibit high lexical diversity, as measured by the Corrected Type-Token Ratio (CTTR) [55], indicating rich semantic content and a robust temporal dimension for evaluation.

4. Towards Temporal Dynamics Evaluation

In this section, we delve into the evaluation of temporal dynamics through a detailed examination of leading Text-to-Video (T2V) models that have been made available as open-source. Our goal is to assess their capabilities in capturing the intricate temporal nuances inherent in video generation.

4.1. Text-to-Video Generative Models

For our analysis, we have selected three T2V models that are not only representative of current technological advancements but are also publicly accessible, ensuring the reproducibility of our findings.

ModelScope [8] is an advanced diffusion-based T2V model, enhancing the Stable Diffusion framework with temporal convolution and attention mechanisms. It's trained on image-text and video-text datasets, demonstrating broad media adaptability.

ZeroScope [42] is a T2V model derived from ModelScope, notable for its watermark-free videos. It's optimized for 16:9 video formats, ideal for widescreen content.

Pika [41] is a publicly available T2V model, that excels at interpreting and stitching together sequential images into fluid video narratives.

Table 2. **Definition of Fine-Grained Evaluation Perspectives for Temporal Dynamics.** The evaluation criteria, derived from our extensive words-bag, cover the most essential aspects of video evaluation, reflecting real-world distributions.

Evaluation Types	Definition	Examples
	Temporal Evaluation Dimensions	
`Explicit event sequence`	For a single entity, two consecutive actions: do A then do B.	*The cat stretches lazily, then curls up for a nap.*
`Implicit event sequence`	For a single entity, two consecutive actions: before, after.	*The rainbow appears after the rain.*
`Explicit event sequence`	or multiple entities executing the same action at a time, two consecutive actions.	*The cars stop at the red light, then accelerate as soon as it turns green.*
`Implicit event sequence`	or multiple entities executing the same action at a time, two consecutive actions: before, after.	*After entering the classroom, students take their seats.*
`Aggregated entities`	One entity at first and then two joins in.	*There are 3 children playing together and then one other child joins in.*
`Scene transition`	One transition, from one scene to the other.	*The scene transitions from inside the café to the street outside.*
`Event timing`	Explicitly points out the time of one action.	*The girl stared at the refrigerator for 2 seconds and then opened the refrigerator.*
`Camera perspective transition`	One transition, e.g. from the side to the front.	*The camera shifts from a high-angle view overlooking the city skyline to a low-angle close-up of a bustling street corner.*
`Direction of movement`	From one direction to the other, e.g. from forward to right.	*The boy runs forward initially, then makes a right turn at the intersection.*
`Emotional change`	From one emotion to another.	*She frowns slightly, then her lips twitch into a faint smile.*
`Shape change`	Transform from one shape to another.	*The leaves wilt, then regenerate, transforming their structure.*
`Weather pattern change`	From one type of weather to another.	*The sun disappears behind clouds.*
`Age change`	Gradually mature.	*The girl grows up from a child and now she is a university student.*
`Acceleration and speed`	Accelerate or Decelerate.	*The athlete sprints faster, then crosses the finish line with lightning speed.*
`Lighting and shadows`	Light changing.	*The storm clouds gather, darkening the sky and deepening the shadows.*
`Cause and effect`	Event changing in appropriate time line due to cause and effect.	*The fire then led to the burning of the house and the family's panicked escape.*

4.2. Experimental Setting

Evaluation Criteria. Our comprehensive benchmarking study evaluates the performance of the models across 16 distinct time scales, utilizing a core subset of 560 samples to enable a thorough analysis of their temporal dynamics. We consolidate our findings into three overarching categories for clarity: Event Dynamics, Visual Dynamics, and Narrative Dynamics. Qualitatively, Event Dynamics evaluates the model's proficiency in understanding and generating video content that accurately reflects the progression of events. Visual Dynamics examines the model's capability to maintain visual consistency and realism throughout the video. Narrative Dynamics, on the other hand, assesses the model's ability to weave coherent and engaging stories, ensuring that the video content is not only visually appealing but also meaningful.

Evaluation Metrics. Our evaluation framework is grounded in human annotations, amassing ratings for each (prompt, video) pair. As delineated in Sec. 3.2, we adhere to stringent guidelines for annotating to ensure fairness and precision in our human ratings. These ratings are scaled from 1 to 5, where 1 indicates the lowest alignment and 5 represents the highest.

4.3. Results and Analysis

Comprehensive T2V Evaluation Leaderboard. Our analysis provides a detailed assessment of T2V models, supported by extensive human ratings to underline their capabilities in handling temporal dynamics. The summarized outcomes, depicted in Fig. 5 alongside detailed comparisons in Tab. 3, shed light on the distinct strengths and areas needing enhancement for each model. ZeroScope leads in the overall text-to-video conversion, followed closely by Pika and ModelScope.

Fine-grained Evaluation across Temporal Dynamics. ZeroScope excels in event dynamics, particularly with Aggregated Entities and Event Timing, indicating its superior handling of complex scenes and precise event sequenc-

Table 3. **Fine-Grained Evaluation Leaderboard on T2VBench.** Elevating beyond existing text-to-video benchmarks, T2VBench offers a detailed assessment of temporal dynamics for contemporary text-to-video generative models. Our analysis reveals that Zeroscope stands out in event dynamics, while Pika showcases exceptional prowess in both visual and narrative dynamics.

Benchmarks	Event Dynamics					Visual Dynamics						Narrative Dynamics	
	Event Sequence	Aggregated Entities	Scene Transition	Event Timing	Cause & Effect	Camera Perspective Transition	Direction of Movement	Geometry Change	Age Change	Acceleration &Speed	Lighting & Shadows	Emotional Change	Weather Pattern Change
ModelScope [8]	3.4	3.6	**3.8**	2.2	**2.7**	3.8	2.9	3.7	3.3	3.2	2.4	2.5	3.3
ZeroScope [42]	**3.6**	**4.0**	**3.8**	**3.1**	2.5	**4.1**	3.1	3.1	2.6	1.8	**4.2**	2.3	**3.9**
Pika [41]	3.2	2.4	3.3	1.6	2.2	**4.1**	**3.6**	**3.8**	**3.7**	**3.3**	3.3	**3.5**	3.7

ing. On the other hand, Pika shines in visual and narrative dynamics, excelling in depicting movement and emotional shifts. This suggests Pika's adeptness at capturing the semantic essence of videos and enhancing environmental storytelling with dynamic elements like changing weather patterns. Nonetheless, Pika falls short in aspects of event dynamics, such as timing and sequence, pointing to a need for improvement in its logical processing capabilities. When compared to ModelScope, Pika resembles a liberal arts scholar, showcasing a nuanced understanding of color and visual dynamics. In contrast, ZeroScope demonstrates exceptional skill in generating scenarios with significant changes, like shifting camera perspectives, while maintaining consistency. This attribute positions it as a promising foundation for AI-driven movie production, indicating its potential to revolutionize content creation with its adept handling of dynamic and complex narrative structures.

Significant Progress, Yet Miles to Go for T2V Generation. Our exploration of three leading open-source text-to-video (T2V) models reveals diverse strengths in T2V generation. Despite this, their performance, as reflected by the annotated scores hovering around a mark of three, denoting *"Has several minor discrepancies"*, underscores the rigorous challenges posed by our comprehensive temporal dynamics benchmark. This outcome signals a pivotal opportunity; we aspire for our evaluation framework to act as a catalyst, empowering T2V models to refine and elevate their video generation capabilities.

5. Automatic Metrics on Temporal Dynamics

To facilitate automatic evaluation, our benchmark introduces insights into automatic metrics and conducts a detailed analysis to determine which metric is best suited for comprehensive text-to-video evaluation.

Automatic Metrics. To ensure a robust and comprehensive evaluation, our benchmark incorporates a diverse array of automatic metrics from different paradigms, including CLIPScore [15], BLIPScore [31], ImageReward [33], PickScore [32], HPSv2 [34], and VQAScore [20], alongside subjective assessments through human preferences. As mentioned in Sec. 2, due to the scarcity of text-to-video models with high evaluation reasoning ability, we select image-based methods for comprehensive evaluation over our benchmark.

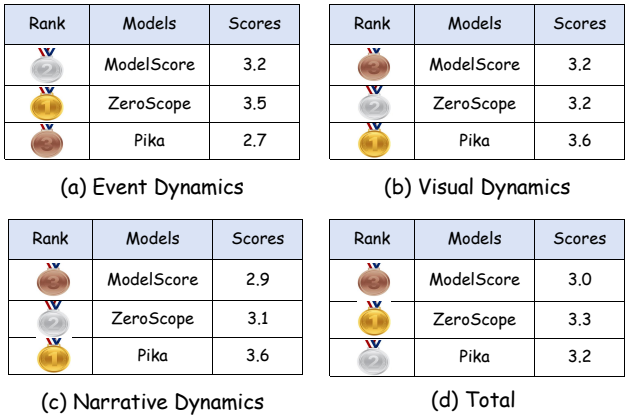

Figure 5. **T2V Evaluation Leaderboard Summary.** ZeroScope emerges as the frontrunner in text-to-video conversion efficacy, with Pika and ModelScope closely trailing. The performance of various models highlights their unique strengths in managing temporal dynamics throughout text-to-video generation.

Experimental Setting. We extract 36 frames from each generated video and feed them individually to the evaluation models, taking the average score as the final score for each sample. Notably, ImageReward scores range from -1 to 1, while all other scores range from 0 to 1.

5.1. Experimental Results

Leaderboard of T2V Evaluation with Automatic Metrics. We report the performance of various metrics on our benchmark in Tab. 4. The results reveal that Zeroscope surpasses other methods when evaluated with advanced metrics, such as VQAScore, which are achieved using Large Foundation Models with an increased number of parameters.

Ambiguity in Metric Preferences. Analysis of results reveals a split in metric preferences: metrics fine-tuned on human feedback show a predisposition towards Pika, whereas VQAScore-based evaluations favor ZeroScope. This divergence indicates the nuanced nature of automatic metrics in capturing different aspects of text-to-

Table 4. **Evaluating Text-to-Video Generation on T2VBench with Automatic Metrics.** We assess text-to-video (T2V) models using established, effective metrics. Our findings indicate significant variability in the discriminative capabilities of these metrics for T2V evaluation. The VLM-based VQAScore exhibits superior evaluative performance across temporal dynamics, whereas the traditional CLIPScore demonstrates limitations in accurately evaluating text-to-video generation.

Metrics	Evaluation Dimensions	Baselines		Finetuned on Human Feedback			VQAScore-based Methods			
		CLIPScore [15]	BLIPScore [31]	ImageReward [33]	PickScore [32]	HPSv2 [34]	InstructBLIP [56]	LLaVA1.5 [57]	CLIP-FlanT5 [20]	ShareGPT4V [58]
ModelScope	Event Dynamics	0.22	0.23	-0.58	0.19	0.23	0.64	0.63	0.64	0.51
	Visual Dynamics	0.23	0.19	-0.68	0.2	0.24	0.68	0.64	0.64	0.54
	Narrative Dynamics	0.2	0.15	-0.62	0.19	0.24	0.65	0.58	0.58	0.48
	Total	**0.23**	0.22	-0.57	0.19	0.24	0.66	0.64	0.64	0.53
ZeroScope	Event Dynamics	0.23	0.29	-0.28	0.20	0.24	0.65	0.70	0.70	0.62
	Visual Dynamics	0.23	0.30	-0.39	0.20	0.24	0.69	0.72	0.72	0.62
	Narrative Dynamics	0.20	0.23	-0.72	0.20	0.24	0.65	0.72	0.72	0.58
	Total	0.22	**0.28**	-0.38	0.20	0.24	**0.67**	**0.71**	**0.71**	0.62
Pika	Event Dynamics	0.22	0.23	-0.29	0.20	0.25	0.63	0.54	0.58	0.66
	Visual Dynamics	0.24	0.34	-0.01	0.21	0.25	0.70	0.63	0.66	0.74
	Narrative Dynamics	0.21	0.24	0.21	0.21	0.25	0.63	0.62	0.62	0.73
	Total	0.22	0.27	**-0.11**	**0.21**	**0.25**	0.66	0.59	0.63	**0.70**

video generation. Further investigation into our diverse and challenging prompts indicates that automatic metrics tend to favor videos with less motion, suggesting a trade-off between visual consistency and semantic richness. Interestingly, this dichotomy aligns with human preferences, which lean towards Pika for its superior handling of visual dynamics and movement, highlighting the complexity of balancing technical precision with the richness of content in evaluating text-to-video models.

5.2. Metric Preferences across Temporal Dimension

Human Preference Alignment. The varying preferences shown by different metrics lead to misalignment and inaccuracies. Therefore, prompted by this ambiguity, we analyze their correlation with human judgment in assessing temporally rich T2V generation. As illustrated in Tab. 5, we explore metrics correlation, exemplified by Pairwise Accuracy [59], Pearson and Kendall [60]. VQA-based methods typically exhibit a higher correlation with human judgments on our benchmark, which is rich in temporal dynamics, benefiting from the generative foundation models. This implies that they could serve as reliable proxies in scenarios where human ratings are scarce. This aligns with the observation that ModelScope outperforms other methods when evaluated against human ratings. Notably, we scale up the Pearson and Kendall with 100 for better demonstration.

Automatic Metrics and Human Rankings. Our analysis presents rankings of automatic metrics against human preference within the context of text-to-video evaluation, as depicted in Tab.5 and Fig.7. It emerges that VQAScore-LLaVA1.5 aligns most closely with human evaluations, demonstrating the highest correlation across Pearson, Kendall, and Pairwise Accuracy metrics. Interestingly, while many metrics appear to diverge from human preferences in text-to-video tasks, VQAScore utilizing the LLaVA backbone stands out by reflecting human judgments more accurately in our benchmark. This observation indicates the importance of selecting appropriate evaluation metrics that resonate with human assessments in the evolving landscape of text-to-video generation.

Table 5. **Metrics Alignment with Human Preferences.** In the challenging text-to-video generation scenarios of T2VBench, both VQAScore and ImageRewards exhibit a stronger alignment with human preferences, suggesting their potential as effective automatic metrics for T2V evaluation.

Methods	Pairwise Acc	Person	Kendall
Baselines			
CLIPScore [15]	42.77 (6)	20.48 (6)	12.31 (6)
BLIPv2Score [31]	39.64 (9)	6.16 (9)	5.07 (9)
Finetuned on human feedback			
ImageReward [33]	50.11 (2)	39.63 (2)	29.27 (2)
PickScore [18]	42.59 (7)	16.41 (8)	11.88 (7)
HPSv2 [34]	42.50 (8)	16.85 (7)	11.67 (8)
VQAScore-based methods			
InstructBLIP [56]	47.46 (5)	29.51 (5)	23.14 (5)
LLaVAv1.5 [61]	**50.42 (1)**	**41.10 (1)**	**29.99 (1)**
ShareGPT4V [58]	49.43 (3)	36.00 (3)	27.69 (3)
CLIP-FlanT5 [20]	47.64 (4)	33.01 (4)	23.56 (4)

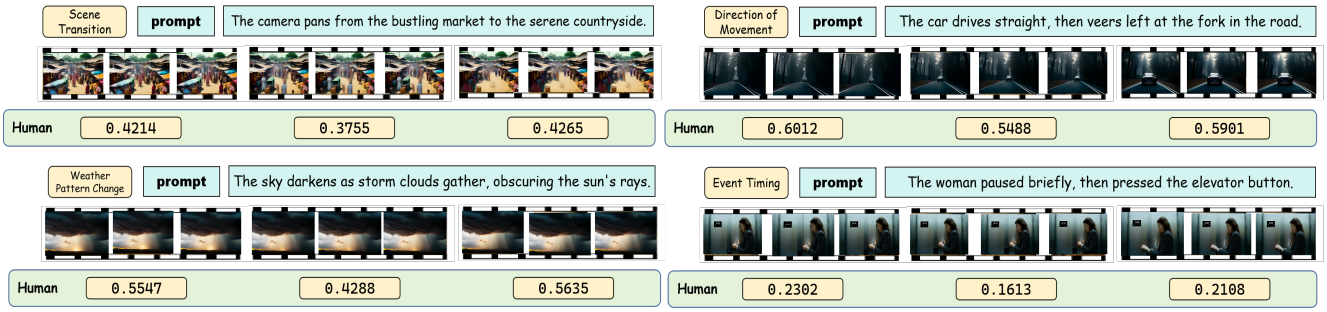

Figure 6. **Temporal Complexities and Challenges Inherent in Video Generation.** We leverage automatic metrics over the entire video duration on some samples from T2VBench, uncovering performance peaks at both the start and end of videos, contrasted by a significant dip mid-video in dynamic contexts. This highlights the intricate temporal challenges faced in video generation.

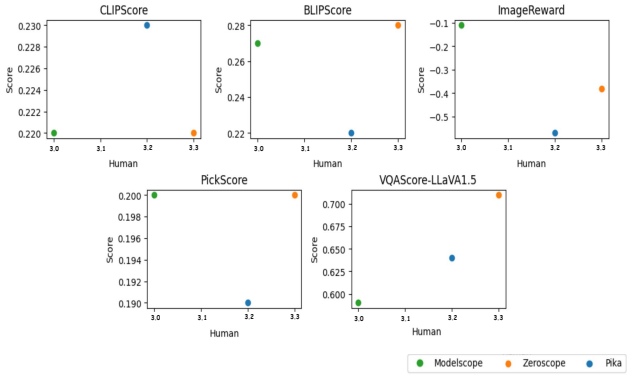

Figure 7. **Correlation Between Automatic Metrics and Human Preferences in Text-to-Video Evaluations.** The VQAScore-LLaVA1.5 demonstrates a superior alignment with human judgments, while T2Vmetric somewhat deviates from human preferences, indicating the critical need for refined automatic metrics in text-to-video evaluations.

5.3. Findings and Insights

Unsuitability of CLIPScore for Text-to-Video Evaluation. Tab. 4 shows that while each metric assesses different facets, CLIPScore cannot distinguish effectively among text-to-video (T2V) outputs, assigning similar scores across varying models and prompts. This highlights its inadequacy for nuanced T2V evaluation. Despite its widespread use, the field must move towards more advanced evaluation methods to capture the complexities of video generation.

Impact of Foundation Model Backbone on T2V Evaluation. The foundation model backbone significantly influences evaluation outcomes within the same framework. As shown in Tab. 4, using InstructBLIP, with low alignment scores, blurs distinctions among models. Conversely, LLaVA1.5, scoring highest in alignment, sharpens differences in our evaluation. This indicates the crucial role of choosing the right foundation model backbone for precise and effective evaluations.

Temporal Dynamics Challenge in T2V Models. Analysis with VQAScore-LLaVAv1.5 reveals a distinct performance pattern in T2V models: high scores at video beginnings and ends, with a dip in the middle, especially in scenarios with significant changes like object movements or scene transitions, as shown in Fig. 6. This pattern indicates the complexity of handling temporal dynamics in video generation, illustrating the unique challenges T2V models face compared to T2I models and affirming the intricate task of generating videos amidst dynamic changes.

Limitations of Current T2V Evaluation Metrics. According to Tab. 5, there's still a significant discrepancy between the top metric's Pearson correlation of 41.10 and human preferences, indicating the inadequacy of existing automatic metrics to reflect nuanced human judgments. However, despite the significant gap between human preferences and automatic metrics, integrating generative models into T2V evaluation processes may enhance evaluation, inspired by the results of metrics that utilize GenFMs.

6. Discussion and Conclusion

Summary. T2VBench marks an important step in text-to-video (T2V) evaluation, emphasizing the crucial role of temporal dynamics. By introducing a comprehensive benchmark with over 1,600 prompts and 5,000 videos rated by humans, this work shines a light on the strengths and challenges of leading models like ZeroScope and Pika in capturing temporal nuances. It underscores the need for advanced evaluation metrics that can better assess temporal coherence in generated videos. Overall, as the pioneering comprehensive text-to-video benchmark, T2VBench not only deepens our comprehension of existing T2V generative models but also provides guidance for future research and development efforts.

Future work. We plan to further explore connections between current image-based automatic metrics and video scenarios involving temporal dynamics, to facilitate the development of generative foundation models.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[2] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023.

[3] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

[4] Mingxiao Huo, Mingyu Ding, Chenfeng Xu, Thomas Tian, Xinghao Zhu, Yao Mu, Lingfeng Sun, Masayoshi Tomizuka, and Wei Zhan. Human-oriented representation learning for robotic manipulation. *arXiv preprint arXiv:2310.03023*, 2023.

[5] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[6] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[7] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

[8] 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.

[9] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[10] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[11] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[12] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[13] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

[14] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[17] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[20] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. 2024.

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[22] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, Weisi Lin, Wynne Hsu, Ying Shan, and Mike Zheng Shou. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*, 2024.

[24] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *CoRR*, abs/2311.16103, 2023.

[25] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. *CoRR*, abs/2311.17982, 2023.

[26] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society, 2016.

[27] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234, 2016.

[28] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019.

[29] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018.

[30] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics, 2021.

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.

[32] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[33] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[34] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *CoRR*, abs/2306.09341, 2023.

[35] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14953–14962. IEEE, 2023.

[36] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11854–11864. IEEE, 2023.

[37] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for step-by-step text-to-image generation and evaluation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[38] 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. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20349–20360. IEEE, 2023.

[39] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[40] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *CoRR*, abs/2311.01361, 2023.

[41] Pika, 2023.

[42] Zeroscope, 2023.

[43] Gen-2: Gen-2: The next step forward for generative ai, 2023.

[44] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *arXiv preprint arXiv: 2311.01813*, 2023.

[45] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost

monocular 3d semantic scene completion in normalized devicecoordinates space. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9421–9431. IEEE Computer Society, 2023.

[46] Jiawei Yao, Tong Wu, and Xiaofeng Zhang. Improving depth gradientcontinuity in transformers: A comparative study on monocular depth estimation with cnn. *arXiv preprint arXiv:2308.08333*, 2023.

[47] Chenwei Lyu, Huai Yu, Zhipeng Zhao, Pengliang Ji, Xiangli Yang, and Wen Yang. Self-supervised dense depth estimation with panoramic image and sparse lidar. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 6819–6822, 2023.

[48] Yutong Wang, Bairan Xiang, Shinan Huang, and Guillaume Sartoretti. Scrimp: Scalable communication for reinforcement- and imitation-learning-based multi-agent pathfinding. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9301–9308, 2023.

[49] Jiongchao Jin, Huanqiang Xu, Pengliang Ji, and Biao Leng. Imc-net: Learning implicit field with corner attention network for 3d shape reconstruction. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1591–1595, 2022.

[50] Pengliang Ji, Angtian Wang, Yi Zhang, Adam Kortylewski, and Alan Yuille. Volumetric neural human for robust pose optimization via analysis-by-synthesis. In *SVRHM 2022 Workshop@ NeurIPS*, 2022.

[51] Jiawei Yao, Xiaochao Pan, Tong Wu, and Xiaofeng Zhang. Building lane-level maps from aerial images. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP)*, pages 3890–3894. IEEE, 2024.

[52] Yizhuo Wang, Yutong Wang, Yuhong Cao, and Guillaume Sartoretti. Spatio-temporal attention network for persistent monitoring of multiple mobile targets. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3903–3910, 2023.

[53] Yutong Wang and Guillaume Sartoretti. Fcmnet: Full communication memory net for team-level cooperation in multi-agent systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 1355–1363, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems.

[54] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14277–14286. IEEE, 2023.

[55] David Malvern and Brian Richards. *Measures of Lexical Richness*. John Wiley Sons, Ltd, 2012.

[56] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[57] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[58] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[59] Daniel Deutsch, George F. Foster, and Markus Freitag. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12914–12929. Association for Computational Linguistics, 2023.

[60] Douglas G. Bonett and Thomas A. Wright. Sample size requrements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65(1):23–28, 2000.

[61] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.