FETV: A Benchmark for Fine-Grained Evaluation of Open-Domain Text-to-Video Generation

Yuanxin Liu[§], **Lei Li**[§], **Shuhuai Ren**[§], **Rundong Gao**[¶], **Shicheng Li**[§], **Sishuo Chen**[¶], **Xu Sun**[§], **Lu Hou**[‡]

[§] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

[¶] Center for Data Science, Peking University [‡] Huawei Noah's Ark Lab

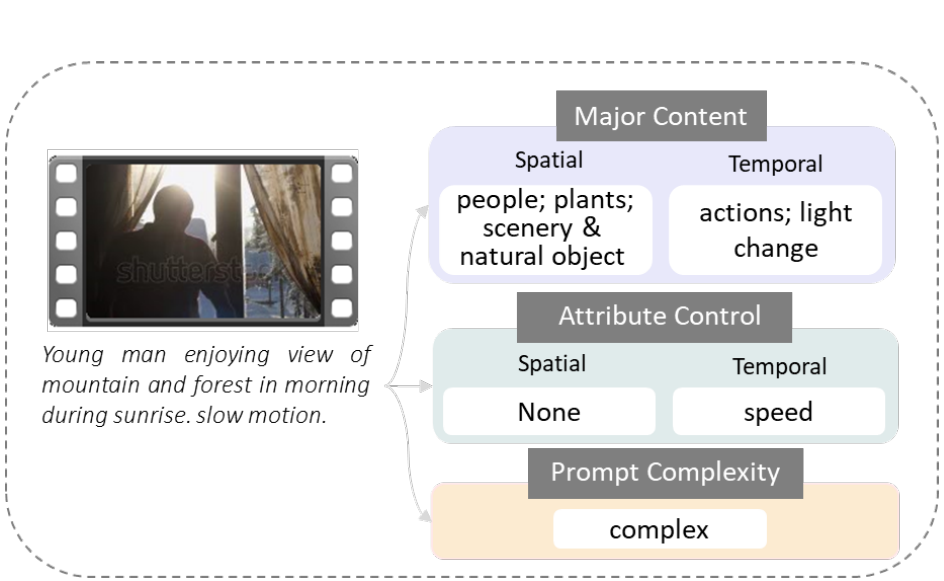
Limitations of Existing Open-Domain T2V Generation Evaluation

- **Lack of fine-grained evaluation**
  - A common practice is to report overall results on entire test set, without considering fine-grained performance on different types of prompts.
- **Lack of reliable automatic metrics**
  - It is unclear whether the automatic metrics are consistent with human standards.

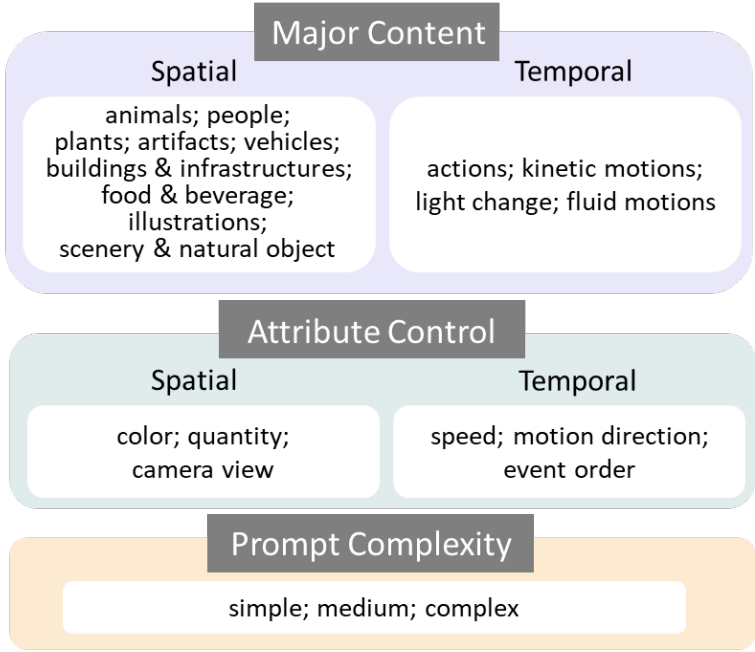
Contributions of This Work

- **Lack of fine-grained evaluation**
  - Propose a benchmark with multi-aspect and temporal-aware categorization.
  - Conduct fine-grained evaluation of representative T2V models.
- **Lack of reliable automatic metrics**
  - Reveal the poor correlation of existing metrics with humans.
  - Present reliable metrics.

Categorization of FETV



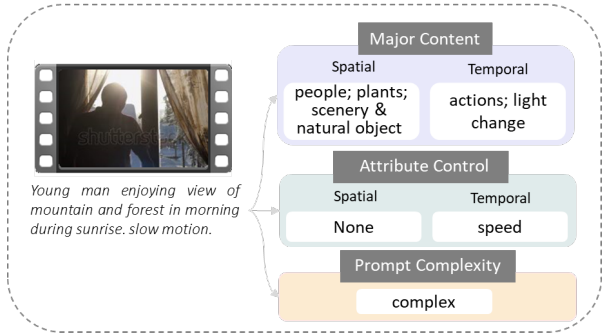
Categorization of Single Data



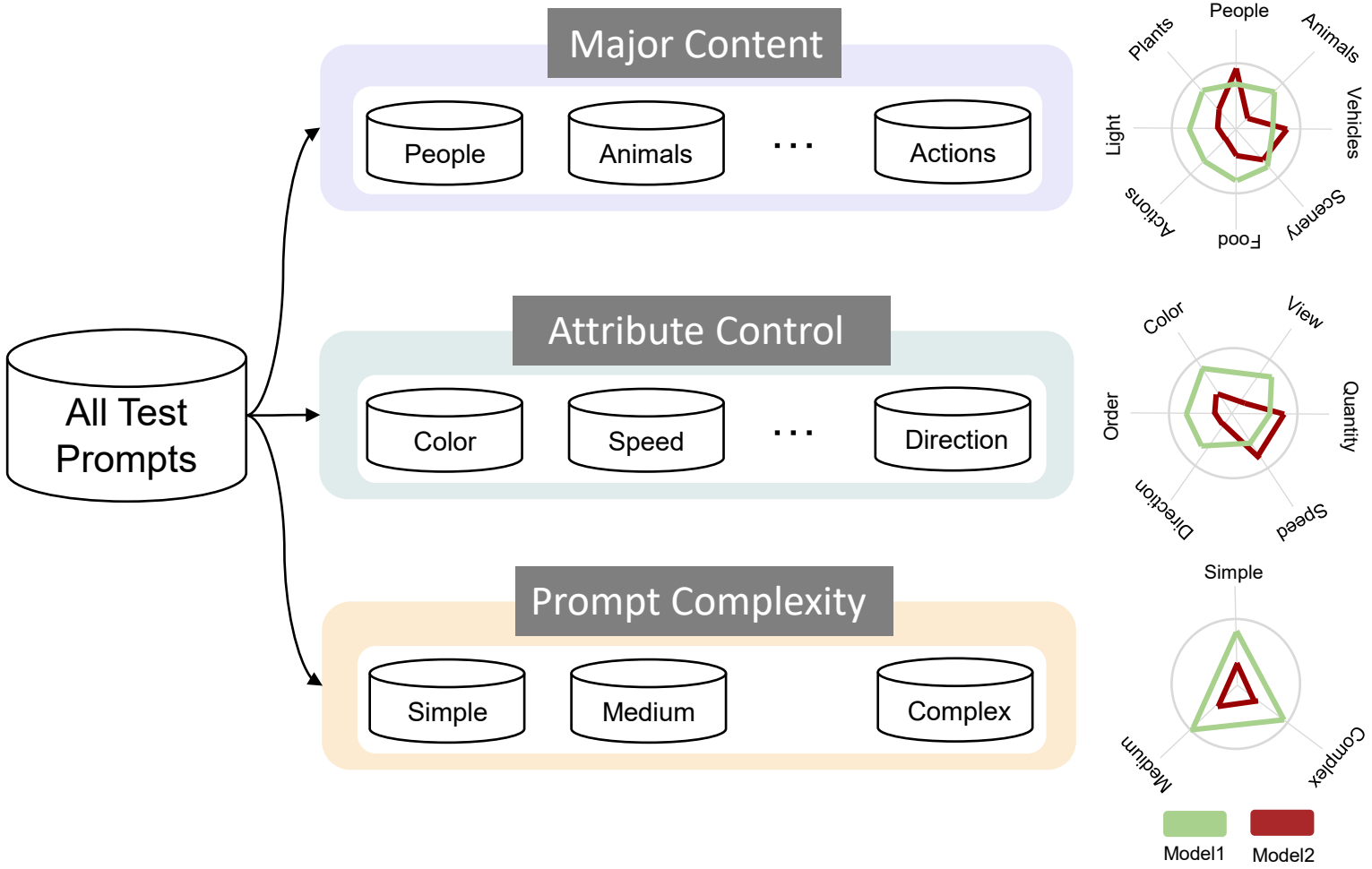
Summary of All Categories

北京大学 PEKING UNIVERSITY

Fine-Grained Evaluation Using FETV



Categorization of Single Data



Advantages of FETV

- **Multi-Aspect**: major content, attribute control, prompt complexity
- **Temporal-Aware**: temporal content, temporal attribute
- **Open-Domain**: diverse prompt categories

Task Type	Benchmark	Open Domain	Major Content		Attribute Control		Prompt Complexity
			spatial	temporal	spatial	temporal	
Text2Image	DrawBench	✓	✗	✗	✓	✗	✗
	PartiPrompts	✓	✓	✗	✓	✗	✓
Text2Video	UCF-101	✗	✗	✓	✗	✗	✗
	Kinetics	✗	✗	✓	✗	✗	✗
	MSR-VTT	✓	✓	✗	✗	✗	✗
	Make-a-Video-Eval	✓	✓	✗	✗	✗	✗
	FETV (Ours)	✓	✓	✓	✓	✓	✓

Data Collection

Prompt Categorization

- Step1: Hand-crafted rules for automatic categorization
- Step2: Manual selection and revision

Prompt Sources

- Prompts of real-world videos from MSR-VTT [1] and WebVid [2]
- Manually written prompts describing unusual scenarios

[1] MSR-VTT: A large video description dataset for bridging video and language.
[2] Frozen in Time: A Joint Video and Image Encoder for End to End Paper.

Data Example

北京大学 PEKING UNIVERSITY

Major Content

Temporal Categories

Actions

Description: Prompts that involve actions of people or animals.

Example Prompts: *people are dancing; a horse is eating*

Kinetic Motions

Description: Prompts that involve the motion of solid objects.

Example Prompts: *the cars drove fast; DNA strands spinning on screen*

Fluid Motions

Description: Prompts that involve motions of fluids or motions like fluids, e.g., the deformation of objects.

Example Prompts: *Time lapse moving clouds with blue sky; the water falls down to the cliff really fast;*

Light Change

Description: Prompts from which the generated videos may involve change of light or color over time.

Example Prompts: *a fire is burning; star time lapse, green aurora moving across the night sky*

Spatial Categories

People

Description: Prompts that involve people.

Example Prompts: *people are dancing; a girl is singing*

Animals

Description: Prompts that involve animals.

Example Prompts: *rabbits are running around; a horse is eating*

Plants

Description: Prompts that involve plants.

Example Prompts: *Coast forest view through water surf; Time-lapse of white hyacinth flowers blooming.*

Artifacts

Description: Prompts that involve human-made objects, not including large-size objects like vehicles.

Example Prompts: *A man is typing on a wireless keyboard. close up.; The musician plays the guitar. close up*

Attributes

Color

Description: Prompts that involve control over color.

Example Prompts: *someone driving an orange sports car; a woman chops a green pepper and an onion*

Camera View

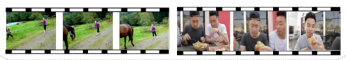
Description: Prompts that involve control over camera view.

Example Prompts: *a first person view of a man driving a red formula one car; overhead view as pingpong players compete on the table*

Quantity

Description: Prompts that involve control over quantity.

Example Prompts: *two young girls playing with a horse; three boys are eating burger;*



Speed

Description: Prompts that involve control over speed.

Example Prompts: *the cars drove fast; bees flying, slow motion shot*

Motion Direction

Description: Prompts that involve control over motion direction.

Example Prompts: *flying counter clockwise around a large yacht; aerial sunrise shot over a town, flying forwards*

Event Order

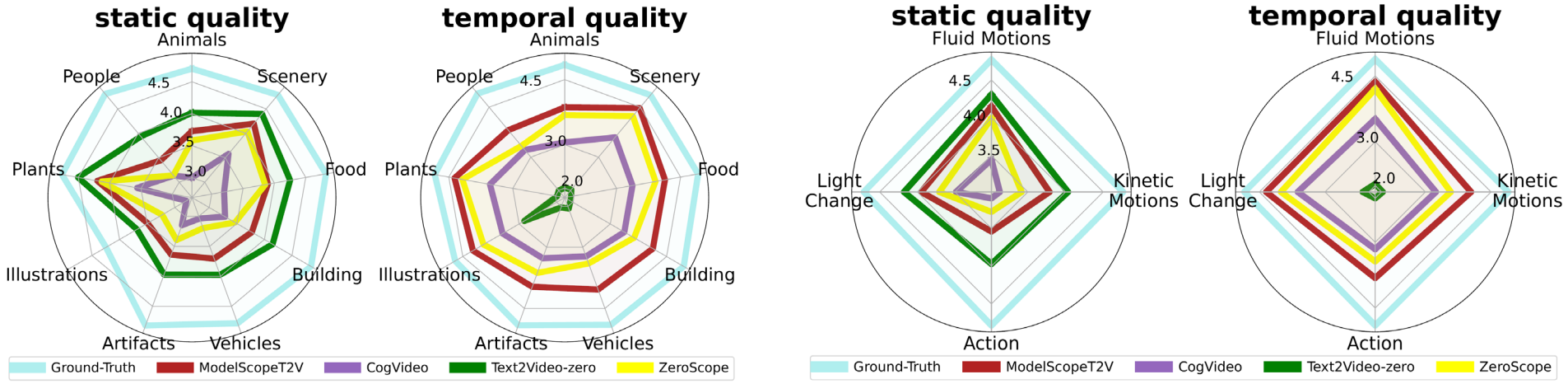
Description: Prompts that involve control over order of events.

Example Prompts: *an old man shakes hands with another man and then they hug each other*

Manual Evaluation: T2V Generation Models

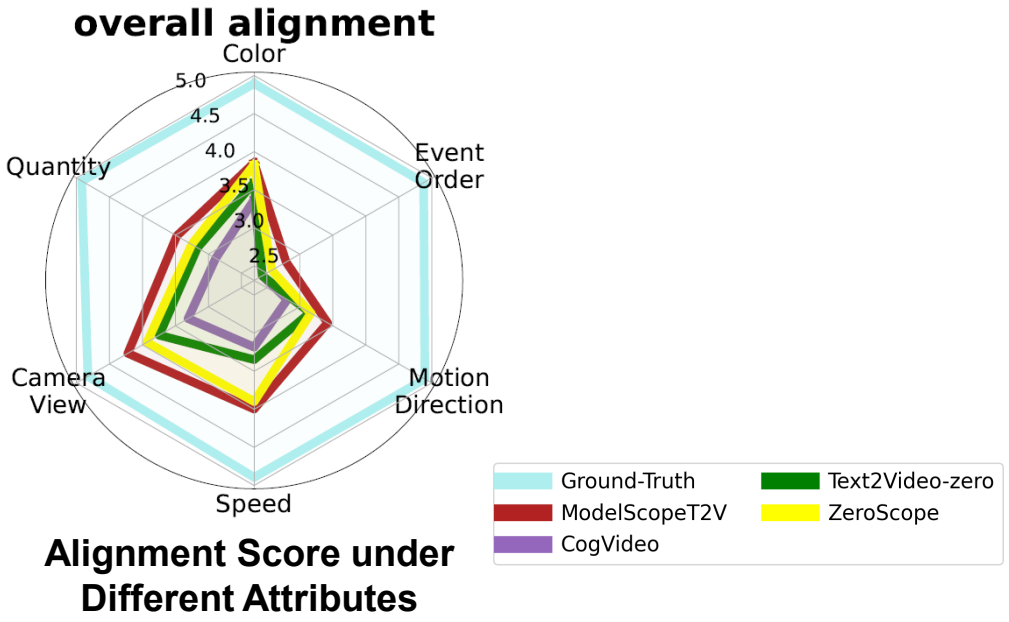
Video Quality

- All T2V model cannot generate ground-truth level videos
- Challenging spatial categories: *People*, *Animals*
- Challenging temporal categories: *Action*, *Kinetic Motions*
- Best static quality: Text2Video-zero, Best temporal quality: ModelScopeT2V

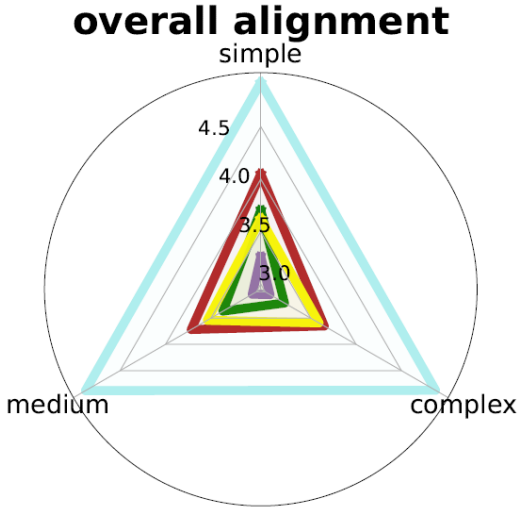


Video-Text Alignment

- T2V models can control *Color* and *Camera View* very well.
- T2V models struggle to precisely control *Quantity*, *Motion Direction* and *Event Order*.
- Videos generated from *Simple* prompts exhibit the strongest alignment, while the difference between *Medium* and *Complex* prompts is not obvious.



Alignment Score under Different Attributes



Alignment Score under Different Prompt Complexity

Diagnosis of Automatic Evaluation Metrics

Video-Text Alignment

- CLIPScore: widely used T2V alignment metric based on the CLIP model
- CLIPScore-ft: Fine-tuning CLIP on video-text retrieval task
- BLIPScore: Replacing CLIP with BLIP
- Otter-VQA (ours): Treating video-text alignment as Video QA using Otter
- UMTScore (ours): Replacing CLIP with UMT

Video Quality

- FID, FVD
- FVD-UMT (ours): Replacing the video encoder in FVD with the UMT model [3]

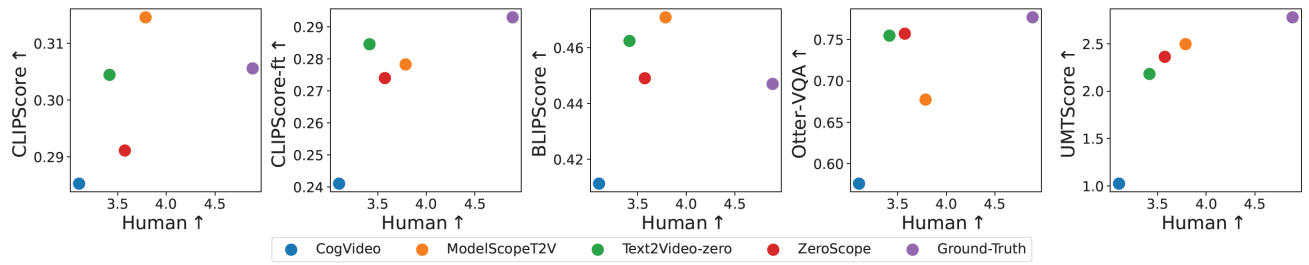
[3] Unmasked teacher: Towards training-efficient video foundation models.

Video-Text Alignment

- The widely-used CLIPScore poorly aligns with humans.
- Fine-tuning CLIP on video-text retrieval is beneficial.
- UMTScore is the only automatic metric consistent with human ranking of T2V models.

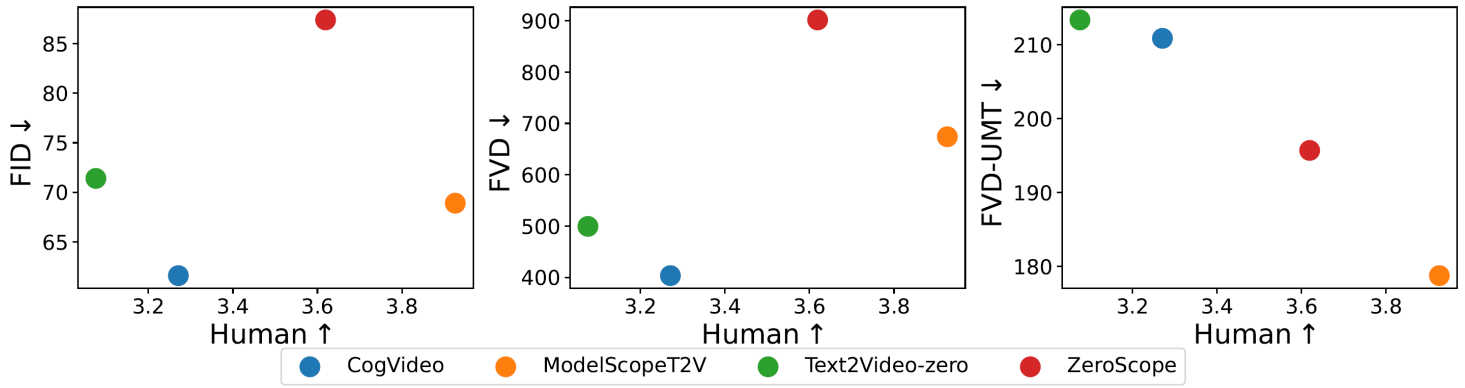
Correlation between automatic and human evaluation of video-text alignment, measured by Spearman and Kendall coefficients.

	Color	Quantity	Camera View	Speed	Motion Direction	Event Order	All
CLIPScore	0.150/0.209	0.165/0.228	0.254/0.345	0.187/0.258	0.153/0.212	0.204/0.284	0.190/0.262
CLIPScore-ft	0.179/0.250	0.309/0.424	0.293/0.402	0.230/0.318	0.283/0.397	0.221/0.313	0.265/0.368
BLIPScore	0.214/0.296	0.227/0.309	0.285/0.394	0.204/0.279	0.238/0.327	0.212/0.292	0.246/0.337
Otter-VQA	0.049/0.070	0.134/0.188	0.027/0.038	0.051/0.073	0.119/0.166	0.146/0.206	0.081/0.114
UMTScore	**0.304/0.420**	**0.394/0.528**	**0.300/0.415**	**0.296/0.407**	**0.356/0.476**	**0.295/0.406**	**0.309/0.425**
Human	0.547/0.702	0.647/0.784	0.447/0.595	0.539/0.683	0.619/.0.747	0.517/0.680	0.576/0.719



Video Quality

- FVD-UMT is the only automatic metric consistent with human ranking of T2V models.



Prompt: a man leading a donkey across a field towards a parked truck

Human	2.0(3rd)	2.0(3rd)	3.0(2nd)	4.67(1st)
UMTScore	-0.965(4th)	0.138(3rd)	2.627(2nd)	4.090(1st)
BLIPScore	0.407(3rd)	0.367(4th)	0.482(1st)	0.478(2nd)
CLIPScore-ft	0.288(4th)	0.304(3rd)	0.333(1st)	0.314(2nd)
CLIPScore	0.301(3rd)	0.306(2nd)	0.337(1st)	0.284(4th)
Otter-VQA	0.767(1st)	0.667(2nd)	0.633(4th)	0.667(2nd)

Thank you!

Benchmark Link: https://github.com/llyx97/FETV