VISTA: A Benchmark for Spatio-Temporal Interaction Understanding in Videos

Supplementary Material

1. Dataset Collection Details

1.1. Query Formulation

A core component of our benchmark is the explicit evaluation of human-style narrative queries ('freeform') versus template-based queries ('referral'). Freeform queries capture open-ended, conversational descriptions, while referral queries focus on concise, object-centric expressions that uniquely identify the primary subject.

Freeform Queries (Q_F): For most datasets (HCSTVG-v1 [3], HCSTVG-v2 [3], VidSTG [4], MeViS [1], RVOS [2]), we use the original captions provided. For VidVRD, which provides relation triplets in the format ⟨subject, predicate, object⟩, we reformulate these into natural language sentences using `gpt-4o`. Table 4 shows the prompt used for this conversion. The conversion preserves all information from the triplet while generating grammatically complete, contextually coherent sentences suitable for natural language grounding tasks.

Referral Queries (Q_R): Referral queries are derived by prompting `gpt-4o` to extract the primary subject phrase that uniquely identifies the subject in the freeform query. For example, given the freeform query *"A man in a suit walks into the room and sits down"*, the referral query becomes *"A man in a suit"*. The extraction focuses on isolating the minimal phrase necessary to uniquely identify the subject, which may include attributes (e.g., "in a suit"), spatial positions (e.g., "to the left"), actions (e.g., "holding a cup"), temporal markers (e.g., "first to arrive"), or simply the entity type when no ambiguity exists. For VidVRD triplets where the subject and object are of the same type (e.g., ⟨red_panda, sit_left, red_panda⟩), we use `gpt-4o` to generate minimal referral phrases directly from the triplets. Table 5 shows the prompt used for VidVRD referral extraction.

1.2. Sample Annotation Pipeline

Our caption annotation pipeline is structured hierarchically, progressing from coarse to fine-grained classification. The pipeline consists of three independent tasks: (1) **Coarse Involved Entity Classification**: categorizing interactions based on the types of entities involved (e.g., Human–Human, Human–Object, Animal–Animal); (2) **Coarse Spatial vs. Temporal Classification**: determining whether the interaction can be inferred from a single frame (spatial) or requires temporal evolution (temporal); and (3) **Fine-Grained Interaction Classification**: assigning one or more appropriate behavioral categories (e.g., Physical, Observation, Cooperative) that capture the nuanced relational dynamics of the interaction.

For each task, we provided `gpt-4o-mini` with a task-specific prompt containing definitions and examples. Table 1 shows the prompt for entity classification, Table 2 shows the prompt for spatial vs. temporal classification, and Table 3 shows the prompt for fine-grained multi-label assignment. The fine-grained interaction axis is explicitly multi-label. A caption often expresses multiple interaction properties simultaneously, for example, aa person may both 'move' and 'observe,' or simultaneously 'manipulate an object' and 'change state.' Limiting this axis to a single label would fail to capture the multi-dimensional nature of visual interactions. Across all datasets, samples have an average **1.73** fine-grained categories, capturing the diversity of interaction types present in a single moment or event.

After annotation, we performed manual review as described in Section 3 of the main paper. This involved iterating through JSON outputs for each dataset, verifying consistency, and adjusting annotations when needed.

2. Evaluation Setup

In this section, we provide further details about our evaluation setup. All models are evaluated in a zero-shot setting using their native prompt formats and bounding box output strategies as specified in their original papers. This ensures consistency with their intended usage and training objectives.
Frame Sampling: All datasets are evaluated using a frame step of 5 (i.e., every fifth frame is processed), except for RVOS, which is pre-sampled at this rate in its original distribution. To generate predictions for intermediate frames, we perform linear interpolation of bounding boxes. For example, given predictions at frame 1 and frame 5, the intermediate frames receive: [frame 1: bbox_1, frame 2: bbox_1, frame 3: bbox_1, frame 4: bbox_5, frame 5: bbox_5].
Video Trimming: Videos are temporally trimmed to include only frames where ground truth annotations are present, ensuring that evaluation focuses on annotated segments.
Resolution: Models use their native input resolutions as specified in their original papers, without forced normalization. We preserve the original video aspect ratios and dimensions.

3. Statistical Significance Analysis

All performance differences reported in the main paper are validated through bootstrap confidence intervals and non-parametric hypothesis tests. For each model, we compute 95% confidence intervals for overall m_vIoU using the percentile bootstrap method ($B = 10{,}000$ resamples, sampling with replacement at the video–query pair level), reporting

Table 1. Prompt provided to `gpt-4o-mini` for Involved Entity Classification.

System:
You are a precise video scene classifier. Classify the caption into one of the following coarse categories:

- Human–Human
- Human–Object
- Human–Animal
- Animal–Animal
- Animal–Object
- Object–Object
- Human–Self
- No Interaction

Few-Shot Examples
- Caption: "The sitting man turns his head to look at the standing man." → Human–Human
- Caption: "The woman sits on a bench." → Human–Object
- Caption: "The person pets the dog." → Human–Animal
- Caption: "Two zebras walk together." → Animal–Animal
- Caption: "The elephant pushes a ball." → Animal–Object
- Caption: "A vase is on top of a table." → Object–Object
- Caption: "The man covers his mouth." → Human–Self
- Caption: "A single person stands alone." → No Interaction

Output Format
Respond with the coarse category only.

Table 2. Prompt provided to `gpt-4o-mini` for Spatial vs. Temporal Classification.

System:
Classify the caption as either:

- **Spatial**: A static visual relationship that can be inferred from one frame.
- **Temporal**: An action, state change, or motion requiring multiple frames.

Few-Shot Examples
- "A man leans on a sofa." → Spatial
- "The man turns his head." → Temporal
- "The person walks toward the door." → Temporal
- "The child stands beside the table." → Spatial

Output Format
Respond with either `Spatial` or `Temporal`.

the 2.5th and 97.5th percentiles as bounds. Results are presented in Table 6. Despite overlapping CIs in some cases (*e.g.*, MiniGPT-v2, Sphinx-v2, and Qwen-VL-Chat), pairwise hypothesis tests confirm that nearly all differences are statistically significant.

3.1. Pairwise Model Comparisons

We conduct two-sided Wilcoxon signed-rank tests on paired per-sample m_vIoU scores for all $\binom{11}{2} = 55$ model pairs, with Holm–Bonferroni correction to control the family-wise error rate at $\alpha = 0.05$. **54 of 55** comparisons yield $p < 0.05$ after correction (Table 7), confirming that nearly every model pair is statistically distinguishable on VISTA. The sole non-significant pair is Qwen-VL-Chat vs. Sphinx-v2 ($p = 0.37$), whose CIs also overlap in Table 6. All other pairs, including the close MiniGPT-v2/Sphinx-v2 and MiniGPT-v2/Qwen-VL-Chat comparisons are significant.

3.2. Referral vs. Freeform Query Structure

Section 4.1 reports that models generally perform better on referral (R) than freeform (F) queries. We validate this with two-sided Mann–Whitney U tests. As shown in Table 8, **10 of 11** models exhibit statistically significant R/F differences. The sole non-significant case is Qwen-VL-Chat ($p = 0.31$), whose R–F gap is negligible ($\Delta = +0.13$). Among significant models, CogVLM shows the largest referral advantage ($\Delta = +10.43$), while MimoVL ($\Delta = +1.21$) exhibits a modest gap. Qwen3-VL reverses the trend with freeform

Table 3. Prompt provided to `gpt-4o-mini` for Fine-Grained Multi-Label Interaction Classification.

System:
You are labeling visual interaction captions into one or more categories.

Categories
Affective, Antagonistic, Body Motion, Communicative, Competitive, Cooperative, Movement, Observation, Passive, Physical Interaction, Provisioning, Proximity, Social, Spatial, Supportive

Instructions
For each caption, return a JSON object with:

```
[
  {
    "caption": "...",
    "categories": ["Movement", "Body Motion"],
    "notes": "Brief reason for classification"
  }
]
```

Multiple categories may be assigned to each caption.
Output Format: Return valid JSON only.

Table 4. Prompt provided to `gpt-4o` for VidVRD Freeform Query Generation from Relation Triplets.

System:
You are an English language expert. Your objective is to generate complete, understandable, cohesive phrases from ⟨subject, predicate, object⟩ triplets. You must make complete sentences without losing any information. You must ensure you always specify the exact same subject and object as per the triplet and never change them. You must ensure you never significantly modify the predicate. You must ensure that the phrase makes complete sense on its own.

Few-Shot Examples
- ⟨person, sit_above, cycle⟩ → "The person sits above a bicycle"
- ⟨red_panda, sit_left, red_panda⟩ → "The red panda sitting to the left of the other red panda"
- ⟨boy, touch, boy⟩ → "The boy who touches the other boy"
- ⟨bear, larger, bear⟩ → "The larger of the two bears"

User:
Generate the phrase from the triplet: ⟨{subject}, {predicate}, {object}⟩

marginally outperforming referral ($\Delta = -1.56$, $p = 0.02$), consistent with our interpretation that broad pretraining enables exploitation of richer freeform context.

3.3. Cross-Entity vs. Same-Entity Gap

Section 4.2 identifies a consistent cross-entity advantage: models ground more accurately when the referent and distractors belong to different semantic classes. We compute mean m_vIoU over cross-entity samples (HA, HO, AO) and same-entity samples (HH, AA, OO) from the per-sample score distribution, then apply Mann–Whitney U tests. As Table 9 shows, the advantage is significant ($p < 0.05$) for **10 of 11 models**, with Qwen-VL-Chat as the sole exception ($p = 0.25$). Gaps range up to 11.6 points (LLaVA-G), confirming that same-entity disambiguation failure is a broadly systematic phenomenon.

3.4. Spatial vs. Temporal Balance

We test for spatial (S) vs. temporal (T) preference using Mann–Whitney U tests per model (Table 10). **9 of 11** models show significant S/T differences, with MiniGPT-v2 ($p = 0.18$) and Sphinx-v2 ($p = 0.09$) as the exceptions. The direction of the bias varies by model family: CogVLM exhibits a strong spatial preference ($\Delta = +11.8$), while MimoVL shows a strong temporal preference ($\Delta = -6.6$). Several specialist models also favor temporal samples, suggesting that superficially balanced aggregate S/T performance masks model-specific directional preferences.

References

[1] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages

Table 5. Prompt provided to `gpt-4o` for VidVRD Referral Query Generation from Relation Triplets.

System:
You are an English language expert. Your objective is to generate complete, understandable, cohesive referral phrases from ⟨subject, predicate, object⟩ triplets. You will receive triplets in which the subject and the object are always the same. You may receive some cases in which the predicate uniquely refers to or identifies the subject and some cases in which it does not. In the first case, you must generate the bare minimum sufficient referral phrase that refers to the subject. In the second case, you must indicate that `is_referral_predicate` *is False.*

Few-Shot Examples
- ⟨red_panda, sit_left, red_panda⟩
`is_referral_predicate`: True → "The red panda sitting to the left."
- ⟨boy, touch, boy⟩
`is_referral_predicate`: True → "The boy who touches the other one."
- ⟨bear, larger, bear⟩
`is_referral_predicate`: True → "The larger bear."
- ⟨girl, next_to, girl⟩
`is_referral_predicate`: False → None
- ⟨zebra, play, zebra⟩
`is_referral_predicate`: False → None

User:
Generate the referral phrase from the triplet: ⟨{subject}, {predicate}, {object}⟩

Table 6. Bootstrap 95% confidence intervals for overall m_vIoU ($n = 10{,}000$ resamples).

Family	Model	m_vIoU	95% CI
Foundation	G-DINO	34.64	[34.08, 35.20]
Generalist	InternVL-2.5	49.73	[49.50, 49.96]
	MiniGPT-v2	45.78	[45.55, 46.01]
	Sphinx-v2	45.82	[45.58, 46.05]
	Qwen-VL-Chat	45.49	[44.94, 46.06]
	Qwen3-VL	63.96	[63.43, 64.49]
	MimoVL	44.54	[44.03, 45.05]
Specialist	Shikra	31.21	[30.82, 31.60]
	Ferret	20.53	[20.17, 20.89]
	CogVLM	54.70	[54.10, 55.28]
	LLaVA-G	27.11	[26.57, 27.65]

Table 7. Selected pairwise Wilcoxon signed-rank tests. Δ = m_vIoU(A) − m_vIoU(B). All p-values are Holm–Bonferroni corrected.

Model A	Model B	Δ	Corrected p	Sig.
Qwen3-VL	CogVLM	$+9.26$	< 0.001	✓
CogVLM	InternVL-2.5	$+4.97$	< 0.001	✓
InternVL-2.5	Sphinx-v2	$+3.91$	< 0.001	✓
Sphinx-v2	MiniGPT-v2	$+0.04$	< 0.001	✓
MiniGPT-v2	Qwen-VL-Chat	$+0.29$	< 0.01	✓
Qwen-VL-Chat	Sphinx-v2	-0.33	$= 0.37$	✗
G-DINO	Shikra	$+3.43$	< 0.001	✓
CogVLM	LLaVA-G	$+27.59$	< 0.001	✓
Qwen3-VL	Ferret	$+43.43$	< 0.001	✓

2694–2703, 2023. 1

[2] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV*, page 208–223, Berlin, Heidelberg, 2020. Springer-Verlag. 1

[3] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:8238–8249, 2020. 1

[4] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10665–10674, 2020. 1

Table 8. Mann–Whitney U test for referral vs. freeform. $\Delta_{\text{R-F}}$ = m_vIoU(R) − m_vIoU(F).

Model	**R**	**F**	$\Delta_{\text{R-F}}$	p**-value**
G-DINO	37.79	32.34	$+5.45$	< 0.001
InternVL-2.5	51.11	48.65	$+2.46$	< 0.001
MiniGPT-v2	46.62	45.13	$+1.49$	< 0.001
Sphinx-v2	47.79	44.28	$+3.51$	< 0.001
Qwen-VL-Chat	45.56	45.43	$+0.13$	$= 0.31$ (n.s.)
Qwen3-VL	62.85	64.41	-1.56	$= 0.02$
MimoVL	43.34	42.13	$+1.21$	< 0.001
Shikra	30.91	31.44	-0.53	$= 0.04$
Ferret	17.74	22.71	-4.97	< 0.001
CogVLM	60.56	50.13	$+10.43$	< 0.001
LLaVA-G	22.51	30.47	-7.96	< 0.001

Table 9. Cross-entity vs. same-entity m_vIoU (per-sample means). Mann–Whitney U tests.

Model	**Cross**	**Same**	Δ	p**-value**
G-DINO	38.54	29.24	$+9.30$	< 0.001
InternVL-2.5	49.51	48.20	$+1.31$	< 0.001
MiniGPT-v2	46.05	44.93	$+1.12$	< 0.01
Sphinx-v2	46.69	45.58	$+1.11$	< 0.01
Qwen-VL-Chat	48.08	47.70	$+0.39$	$= 0.25$ (n.s.)
Qwen3-VL	63.96	63.32	$+0.65$	< 0.001
MimoVL	37.17	42.18	-5.01	< 0.001
Shikra	33.87	31.57	$+2.31$	< 0.001
Ferret	23.44	24.40	-0.97	< 0.001
CogVLM	55.11	46.61	$+8.49$	< 0.001
LLaVA-G	39.01	27.41	$+11.60$	< 0.001

Table 10. Spatial vs. temporal m_vIoU. Mann–Whitney U tests. $\Delta_{\text{S-T}}$ = m_vIoU(S) − m_vIoU(T).

Model	**S**	**T**	$\Delta_{\text{S-T}}$	p**-value**
G-DINO	35.0	30.8	$+4.2$	< 0.001
InternVL-2.5	46.3	48.0	-1.7	$= 0.048$
MiniGPT-v2	43.1	44.3	-1.2	$= 0.18$ (n.s.)
Sphinx-v2	42.6	45.0	-2.4	$= 0.09$ (n.s.)
Qwen-VL-Chat	45.7	45.3	$+0.4$	< 0.001
Qwen3-VL	64.8	64.3	$+0.5$	< 0.001
MimoVL	36.9	43.5	-6.6	< 0.001
Shikra	29.9	32.4	-2.5	< 0.001
Ferret	20.9	23.8	-2.9	< 0.001
CogVLM	57.5	45.7	$+11.8$	< 0.001
LLaVA-G	28.1	31.9	-3.8	< 0.001