

AnyCap: Omni-Modal Captioning with Instruction Alignment

Yiming Ren^{1,2*} Zhiqiang Lin^{1*} Yu Li^{1*} Gao Meng^{1*} Junjie Wang¹
Zicheng Lin¹ Jifeng Dai^{1,2} Yujiu Yang¹ Wenhai Wang^{2,3†} Ruihang Chu^{1†}

¹Tsinghua University ²Shanghai AI Laboratory ³The Chinese University of Hong Kong

rym24@mails.tsinghua.edu.cn, ruihangchu@mail.tsinghua.edu.cn, wangwenhai@pjlab.org.cn

<https://github.com/qishisuren123/AnyCap>

Abstract

We present AnyCap, a plug-and-play framework that brings instruction alignment to omni-modal captioning. Captions offer a unified language interface for multimodal learning, and users increasingly expect instruction-driven control over their content and style. Current caption models lack explicit instruction supervision and are weak at instruction following, while directly tuning them can degrade general language ability. Achieving instruction alignment in an omni-modal setting is harder still, as each modality calls for separate models and custom designs. To address these challenges, AnyCap leverages a multimodal revision-based refinement paradigm that refines uncontrolled captions from existing models to instruction-aligned ones, without re-training base models. By processing multi-modality features in a unified framework, it enables one model to serve images, videos, and audio. To address the lack of instruction-based data, we construct AnyCapData, a large-scale, high-quality corpus spanning three modalities with 28 well-designed instruction types, each with multiple instruction samples. For evaluation, we address the limitations of current metrics for instruction-oriented captioning by designing AnyCapEval. Its key insight is to decouple evaluation into content and style for fine-grained assessment. Extensive experiments show that on AnyCapEval and diverse public benchmarks, AnyCap consistently improves both caption quality and instruction adherence for both open-source and API-based models. Notably, AnyCap-8B boosts GPT-4o’s content scores by 46% and style scores by 12%. Our code and models will be made publicly available.

1 Introduction

Multimodal large language models (MLLMs) are rapidly advancing towards omni-modal intelli-

gence (Hurst et al., 2024; Anil et al., 2023; Xu et al., 2025; Zhao et al., 2025). Textual captions provide a natural way to connect different modalities by conveying their information in a unified language form. Beyond producing generic descriptions, users increasingly expect captions that are instruction-aligned, directly adapting to desired content and style. This can more flexibly support multimodal tasks such as retrieval (Radford et al., 2021; Ji et al., 2023), question answering (Li et al., 2022), and content generation (Ramesh et al., 2022).

Yet, instruction-aligned captioning faces three significant challenges. (i) Existing models have limited instruction following ability. Many are trained as general-purpose understanding models. They can lean toward the requested focus, but mainly produce broad content or unrelated details (Wang et al., 2025b). For finer control over captions, current approaches usually rely on special tokens (Dwivedi et al., 2024) or bounding boxes (Zhao et al., 2024; Huang et al., 2024), to guide local descriptions. In practice, however, users could prefer flexible natural language instructions (*e.g.*, in Table 1) to specify caption types, focuses, and styles. This represents a more fine-grained and natural form of control, yet current captioning models fall short in this ability. Bridging the gap is non-trivial, as directly fine-tuning these models to emphasize such instructions can weaken their overall language capability (Luo et al., 2023). (ii) Adapting instruction alignment to different modalities further adds complexity. Each modality imposes distinct demands. For example, visual captions often emphasize object appearance, whereas audio captions may reflect prosodic cues (Sharma et al., 2023; Mei et al., 2022). Meeting heterogeneous requirements usually forces the use of separate models. (iii) High-quality training data and benchmarks for omni-modal captioning are also lacking. Most caption datasets cover a single modality (Wang et al., 2024c; Qian et al., 2025). They also provide only

*Equal contribution.

†Corresponding Author.

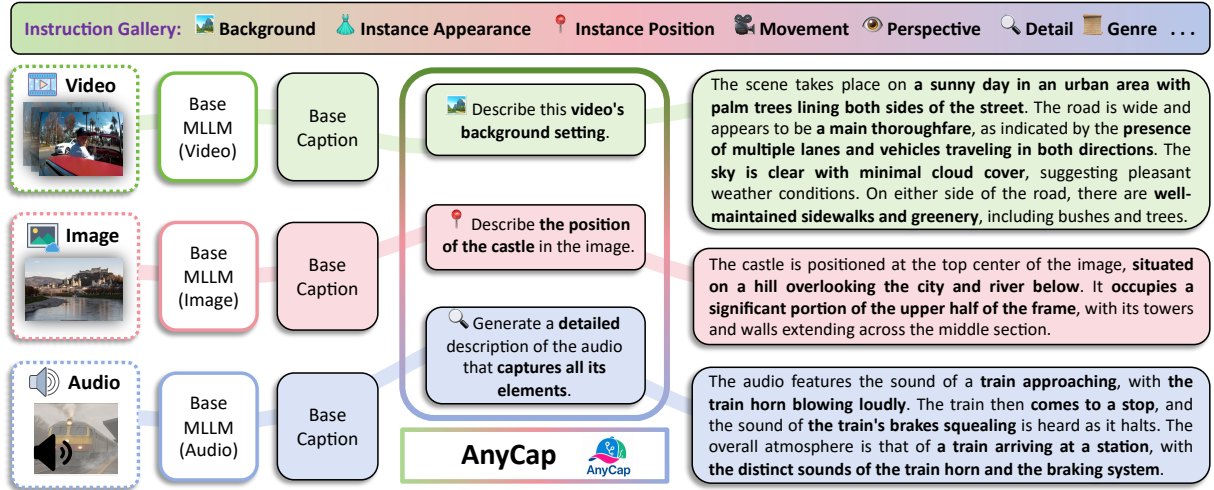


Figure 1: We design diverse natural language instructions for controllable captioning. Given image, video, and audio inputs, base models first generate initial captions. AnyCap effectively aligns them into high-quality, instruction-aligned outputs. A single AnyCap model serves all three modalities.

coarse-grained captions, which cannot support the fine-grained instruction following required above. In particular, audio data is especially scarce, with existing corpora offering only short captions and few long-form descriptions (Kim et al., 2019).

To tackle these challenges, we propose AnyCap, a new captioning framework characterized by two key features. It (i) generates captions that are directly controlled by user language instructions, and (ii) a single model can handle images, videos, and audio data. Inspired by Aligner (Ji et al., 2024), we reduce the task difficulty by shifting the focus from direct generation to a revision-based refinement paradigm. Specifically, we leverage open-source models and available APIs to provide basic captions, and then learn how to align them with user instructions using both the modality observations and instruction texts (Fig. 1). This design simplifies the task, where the base caption supplies general content, while the learning process can focus on capturing instruction intent and refining correctness. Once trained, AnyCap can be used as a plug-and-play module to stack upon diverse captioning models, enhancing captioning without modifying their parameters, or as a flexible tool for re-annotating datasets. Different from Aligner’s static text-to-text refinement, AnyCap performs instruction-aligned caption refinement by integrating three sources of information: multimodal input, coarse caption content, and user instruction. As a result, the model can both preserve/correct factual semantics and follow diverse control intents in an omni-modal setting. Moreover, while Aligner corrects text toward a single fixed alignment objective, AnyCap is

inherently omni-modal and its alignment target is dynamically specified by each user instruction q , rather than fixed in advance.

For the lack of instruction-based caption data, we construct a large-scale dataset termed AnyCap-Data. Unlike prior datasets where instruction types are fixed (e.g., “describe ... in short/details” (Onoe et al., 2024)), we carefully design 28 diverse instruction dimensions spanning images, videos, and audio to better reflect real-world needs (Table 1). To support our alignment paradigm, each sample is annotated as a triplet: a data-aware instruction tailored to the individual content, a preferred caption, and a less-preferred caption, clearly capturing differences in instruction alignment quality. These triplets are generated in a data-aware manner by prompting large MLLMs with the original modality inputs, ensuring that user instructions are grounded in the source content. To further secure quality, we perform small-scale manual screening for each instruction pattern and modality before large-scale generation. With this limited human effort, we improve the quality of automated data generation.

For reliable evaluation, existing methods also remain inadequate. They typically adopt traditional machine translation scores (e.g., BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015)) that lack semantic awareness, or rely on direct LLM-based scoring, which suffers from high randomness across instruction types (Qian et al., 2025; Chen et al., 2025b). We thus design AnyCapEval, with the key insight to categorize diverse instruction types into two orthogonal dimensions, *i.e.*, content and style, and evaluate them separately for focused

and fine-grained evaluation. For content evaluation, we introduce the Keypoint Density (KPD) metric. It employs an automatic matcher to measure the recall of content keypoints and normalizes for caption length. For style evaluation, we rigorously design detailed scoring rubrics for each instruction type, combined with references to guide evaluation and reduce scoring variance. This structured design enables consistent distinction between compliant and non-compliant stylistic outputs.

Extensive experiments show that AnyCap significantly improves caption quality and instruction-following ability across image, video, and audio. It benefits both open-source and proprietary API-based models. Specifically, stacking a 2B AnyCap on GPT-4o improves content following by 35% and style controllability by 8%. The improvements are also consistently shown on widely-used benchmarks including MIA-Bench and VidCapBench, demonstrating its generalizability. We further validate AnyCap through detailed ablation studies and user studies, observing that, compared to other models, AnyCap tends to convey the instruction-required information more accurately while removing redundant content (Table 12), thereby better matching user preferences.

2 Related Work

Image, video, and audio captioning. Most early captioning models focus on a single-modality and produce only brief sentences (Vinyals et al., 2015; Venugopalan et al., 2015; Kim et al., 2019). Later research on image and video captioning moves toward longer and finer-grained descriptions (Celona et al., 2023; Chen et al., 2024a; Ge et al., 2024; Krishna et al., 2017a). Recently, some base MLLMs have attempted to unify image and video captioning (Chen et al., 2024c; Bai et al., 2025) and even cover audio (Hurst et al., 2024; Anil et al., 2023; Xu et al., 2025). Yet, under open-ended prompts these models often drift from user-specific needs, producing irrelevant details or hallucinations that misalign with the intended caption instructions (Han et al., 2024; Huang et al., 2023); this instruction-following deficiency is even more pronounced for speech and audio inputs (Wang et al., 2025a). To address this, we introduce a plug-and-play model that aligns coarse captions across image, video, and audio into fine-grained, high-utility outputs.

Fine-grained captioning with instructions. A range of works use control signals as instructions

to specify caption content or style. Representative approaches include employing length embedding for controlling caption length (Zeng et al., 2023), bounding boxes (Cornia et al., 2019) or masks (Wang et al., 2023) for region-level grounding, soft prompting for style or domain transfer (Zhao et al., 2024), and structured prompts to standardize instance-level descriptions (Fan et al., 2024). While effective, these control interfaces are individually designed and difficult to compose, limiting their ability to capture diverse and complex fine-grained requirements. In contrast, we focus on flexible natural language as a unified instruction interface. Our framework aligns captions with 28 instruction types across modalities to broadly cover the spectrum of captioning needs.

Caption evaluation metrics. Traditional word-overlap metrics like BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015) are widely used, but they cannot evaluate semantic consistency or instruction-following ability. This limitation arises because n-gram similarity only measures surface word overlap. Another common approach is to use LLM or MLLM as judges (Maeda et al., 2024; Qian et al., 2025; Chan et al., 2023; Sun et al., 2024; Mao et al., 2024; Dong et al., 2024; Lu et al., 2024; Liu et al., 2023; Sun et al., 2025), which offers semantic coverage but suffers from large variance and model bias (He et al., 2024; Qiu et al., 2023). Recent attempts also tailor metrics toward specificity in long captions (Chen et al., 2025a), or anchor scoring to verifiable evidence for multimodal understanding (Ren et al., 2026). We incorporate the strengths of these directions by introducing the AnyCapEval benchmark. It explicitly separates *content* and *style* assessment with a particular focus on instruction-following evaluation. We design a new keypoint density (KPD) metric and detailed scoring rubrics for more reliable assessment of caption quality.

3 AnyCap

3.1 Framework

Given a modality input m from images (m^{img}), videos (m^{vid}), or audio (m^{aud}), captioning across different modalities often relies on separate MLLMs, and the resulting captions often fail to follow user-specified content or style requirements. To address this, we design AnyCap as a plug-and-play approach to help existing foundation models align captions with user instructions. As illustrated

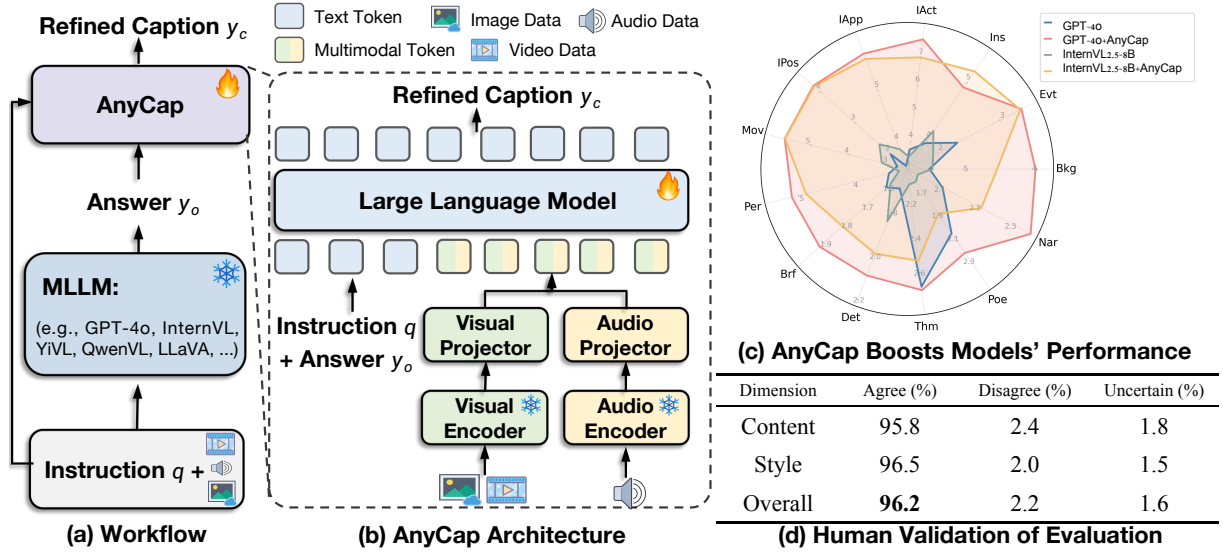


Figure 2: **AnyCap framework.** (a) AnyCap is stacked upon various MLLMs, refining their initial captions into high-quality, instruction-aligned outputs. (b) Specifically, AnyCap takes as input the initial caption, the original modality data, and the user instruction to produce the final caption. (c) Integrating AnyCap consistently boosts base models across both content and style dimensions. (d) AnyCapEval judgments highly align with human preference.

in Fig. 2(a), a frozen base model first processes m to produce an initial caption y_0 . Then, AnyCap ($\mathcal{M}_a$) takes the original input m , the initial caption y_0 , and a detailed user instruction q to generate a refined caption y_c , formulated as

$$y_c = \mathcal{M}_a(m, q, y_0). \quad (1)$$

Concretely, as shown in Fig. 2(b), AnyCap first employs modality-specific encoders (e.g., InternViT (Chen et al., 2024c) for m^{img} and m^{vid} , and EAT (Chen et al., 2024b) for m^{aud}) to extract features from each modality input. These modality features are projected into a shared semantic space via modality-specific linear transformations (MLPs). Concurrently, q and y_0 are tokenized and embedded into textual embeddings. Finally, modality and textual embeddings are concatenated and fed into a large language model to produce the refined, instruction-compliant caption y_c .

The architecture of AnyCap allows us to adopt a multimodal revision-based refinement strategy. Although supervision still follows standard next-token prediction on the final caption y_c , AnyCap focuses on learning the revision pattern from the initial caption y_0 to the instruction-compliant target. This process seamlessly integrates three conditions (*i.e.*, the raw multimodal features, the initial caption semantics, and the specific user instruction), ensuring that the final caption remains grounded in the source input while aligning with the specified requirements. With y_0 as reference, we observe that

revision-based refinement makes the task easier and delivers two gains: (i) it outperforms directly fine-tuning base MLLMs with instructions alone, as shown in Tables 9 and 10; and (ii) it enables a single model to be trained jointly across image, video, and audio data. We only train the lightweight plug-and-play AnyCap, leaving the base MLLMs frozen to preserve their original capabilities. To help the model recognize when *no* change is needed, about 40% of training instances already satisfy the instruction. It makes training more stable and mitigates the over-editing issue, as ablated in Fig. 5. Empirically, AnyCap consistently improves instruction alignment across backbones (Fig. 2(c)), and our automatic evaluation shows strong agreement with human judgments (Fig. 2(d)).

3.2 AnyCapData

We propose AnyCapData that features three aspects: (i) it covers large-scale, high-quality caption data across images, videos, and audio; (ii) each caption is paired with user instructions to specify data-aware requirements; and (iii) each sample adopts a triplet structure to support our multimodal revision-based refinement training, formulated as (q, c, a) . Here, q denotes the language instruction, c is a high-quality caption adhering to the instruction, while a is a suboptimal caption that may exhibit minor deficiencies in factual accuracy, level of detail, or compliance with the instruction.

Table 1: Supported instruction types across image, video, and audio modalities in AnyCapData. Checkmarks (✓) indicate where each control is applicable.

Type	Description	I	V	A
<i>Content</i>				
Bkg	Provide or suppress scene-background details		✓	
Evt	Require mention of an event or temporal change			✓
Ins	Emphasise or ignore specific entities	✓	✓	
IAct	Describe the motion state of a designated instance		✓	
IApp	Characterise visual attributes of a designated instance	✓	✓	
IPos	Specify spatial position of an instance within the scene	✓	✓	
Mov	Specify camera motion type (e.g., pan, zoom, dolly)		✓	
Per	Require a particular viewpoint of an object/person	✓	✓	
Reg	Restrict description to a specified image/video region	✓	✓	
<i>Style</i>				
Brf	Produce a concise rendition with minimal elaboration	✓	✓	✓
Det	Regulate the required level of descriptive granularity	✓	✓	✓
Gen	Adopt a literary form (e.g., poem (Poe), narrative (Nar))	✓	✓	✓
Len	Constrain caption size (words or sentences)	✓	✓	✓
Thm	Conform to a designated linguistic style	✓	✓	

Instruction types. To ensure both diversity and practicality, we identify instruction types by systematically surveying caption-related literature to surface common control categories and by analyzing typical downstream requirements. The final selections are further conditioned on the distinctive characteristics of each modality, as listed in Table 1. More details are illustrated in Appendix A.2.

Construction pipeline. With the designed instruction types, we first use MLLMs to jointly generate a specific user instruction q and its compliant caption c for each image, video, or audio sample. This avoids cases where a separately predicted instruction might request information the caption cannot provide. The prompts for data construction are tailored per modality and instruction type, including clear task descriptions, explicit length and style constraints, common pitfalls, few-shot exemplars, and the original media as input. For audio, we include the dataset’s reference caption (when available) in the prompt to reduce hallucination. Importantly, q is instantiated for each example (e.g., “describe the watermelon’s appearance in this image”, rather than a generic “describe details”), enabling fine-grained and targeted control. We primarily adopt strong open-source VLMs (e.g., InternVL2.5-78B, Qwen2.5-VL-72B) for visual modality. GPT-4o is selected for harder controls (e.g., perspective, poetry) and for audio modality, balancing coverage and cost. For contrastive supervision, given (q, c) we derive a suboptimal caption a by removing guidance to make it uncontrolled or injecting degradations to increase hallucination. The latter includes omitting required keypoints, introducing mild factual errors, or violating style/length constraints in the output.

Before large-scale automatic construction, each



Content

Instruction: Describe the drawing of a face on the ground in the image. The generated caption’s length needs to be 30 to 120 words.

Key points:
describe the simplicity of the face describe the shape of the face
describe the mouth mention the material used describe the eyes
describe ear shapes mention the written word below the face
mention the overall style

Caption: The image shows a face drawn on the ground, likely using chalk or a similar medium. The face is simple yet expressive, featuring exaggerated features like large eyes and a wide mouth. The lines are bold, with clear outlines, giving it a playful, cartoon-like appearance.

Grading:
describe the simplicity of the face: 1/1
mention the written word below the face: 0/1
mention the material used: 1/1
describe the mouth: 1/1
describe the eyes: 1/1
describe ear shapes: 0/1
describe the shape of the face: 0/1
mention the overall style: 1/1



Style

Instruction: The generated caption’s length needs to be 10 to 20 words. Provide a brief caption that accurately captures the key elements of this image. The generated content needs to truly reflect the content of the image, please do not add non-existent content.

Reference: A stone fortress overlooks a vast blue sea with distant mountains and fields in the background.

Scoring Rules: (0-4)
0: Severe quality issues or full hallucination
1: Significant quality issues or major hallucination
2: Slightly inferior to reference or limited hallucination
3: Comparable to reference and no hallucination
4: Slightly better than reference and no hallucination

InternVL2.5-8B response: Stone fortress overlooking vast ocean and distant coastline under clear blue skies.
Score: 2 Reason: Model output captures the core elements like ‘stone fortress’ and ‘vast ocean’ but misses the ‘mountains and fields’ visible in the image.
GPT-4o response: A stone fortress overlooks a vast blue sea with distant mountains and greenery in the background under a clear sky.
Score: 3 Reason: The model output is factually aligned with the image and reference, adequately describing the scene without any hallucination.

Figure 3: **Details of AnyCapEval.** Examples for content evaluation via KPD and style evaluation via rules.

prompt template is manually validated on about 20 instances per modality–type. The validation requires full instruction compliance, no hallucination, correct formatting, and preservation of modality-specific details. After large-scale generation, we manually verify about 5% of the data, achieving over 95% agreement with human preferences (annotation and train/test protocols in Appendix C.5). We also apply type-specific length gate and automatic sanity check to filter degenerate or repetitive outputs. The entire pipeline requires only limited human intervention to enable low-cost scaling.

Statistics. AnyCapData contains $\sim 300k$ (q, c, a) triplets across images (125k), videos (100k), and audio (75k). These original samples are collected from multiple diverse public datasets (Wang et al., 2024b; Onoe et al., 2024; Urbanek et al., 2024; Cui et al., 2024; Fan et al., 2024; Kim et al., 2019) with ensured quality. Data sources and splits, detailed construction procedures, and comprehensive statistic details are shown in Appendix A.

3.3 AnyCapEval

We design AnyCapEval for reliable evaluation of caption quality and, more importantly, the degree of instruction alignment. Beyond previous works, our key idea is that diverse instructions can be grouped into two dimensions: content or style requirements. Accordingly, the evaluation targets: (i) *Content*, which checks whether captions follow the

requested control and remain semantically relevant; and (ii) *Style*, which measures consistency with human references in narrative manner, structure, and fluency. They are separately evaluated to prevent mutual conflicts, for example when polished language conceals factual mistakes.

Content evaluation. Given a ground-truth caption c , a predicted caption y_c , and a user instruction q , we first annotate a keypoint set $\mathcal{K}_{r,q} = \{k_1, k_2, \dots, k_n\}$ covering the information required by q . GPT-4o then serves as an automatic matcher to identify which subset of keypoints is present in y_c (see Fig. 3(a) for examples). Crucially, we decouple semantic understanding from scoring by reframing evaluation as a constrained binary classification task. This leverages LLM reasoning while eliminating generative variance, ensuring a deterministic and reproducible process.

To address the verbosity issue, where longer captions may trivially cover more keypoints, we introduce the keypoint density (KPD) to normalize for length, formulated as $\text{KPD}(y_c; q) = N_{\text{match}}/L_{\text{words}}(y_c) \times 100$. N_{match} is the number of matched keypoints and $L_{\text{words}}(y_c)$ is the word count of y_c . KPD quantifies the effective information rate, rewarding concise captions that meet the instructions while penalizing redundant or irrelevant content; see Appendix B for full details.

Style evaluation. We employ GPT-4o to compare the predicted caption y_c with the ground truth c under instruction q , producing a discrete score $s(y_c) \in \{0, \dots, 4\}$. The scoring criteria are as follows: (0) severely deviates from q or largely hallucinated/incorrect; (1) deviates from q or contains many hallucinations; (2) slightly deviates from q or contains certain hallucination; (3) highly similar to c and hallucination-free; and (4) outperforms c while remaining hallucination-free (see Fig. 3(a)).

This scoring follows explicit criteria along three aspects: semantic similarity, hallucination severity, and stylistic conformity, for each level from 0 to 4. By restricting evaluation to a deterministic matching task, the evaluation reduces the judge’s variance for improving reproducibility.

To verify AnyCapEval’s reliability, we sample 200 examples per model and ask five annotators to judge whether predicted scores and rationales (e.g., Fig. 3(a)) match human preference. The results in Fig. 3(b) show strong alignment between AnyCapEval and human judgments (Spearman $\rho \geq 0.88$). We further validate that replacing the GPT-4o judge with an alternative model yields con-

sistent rankings, and that scores remain stable under judge-prompt perturbations; full results are in Appendix D. We also observe that integrating AnyCap consistently improved strong base models in both content and style dimensions (Fig. 2(c)).

4 Experiments

We train AnyCap in two variants (2B and 8B) on our proposed AnyCapData. Because AnyCap is a relatively small plug-and-play model, it only requires 6 hours on 32 NVIDIA A100 (80GB) GPUs for the 2B variant and 21 hours for the 8B variant to converge. Implementation and training details are provided in Appendix C.1; with a detailed hyperparameter summary in Table 19. Both variants are initialized from InternLM (Cai et al., 2024) with all modality encoder backbones kept frozen throughout training.

4.1 Evaluation on Instruction-Aligned Captioning

We evaluate the impact of integrating AnyCap with various powerful base models (proprietary models, e.g., GPT-4o (Hurst et al., 2024); open-source models, e.g., InternVL2.5 (Chen et al., 2024c), Qwen2.5-VL (Bai et al., 2025), MiniCPM-o (Yao et al., 2024)) using our AnyCapEval benchmark across the image, video, and audio modalities.

As shown in Tables 2 to 4, adding AnyCap results in significant improvements in instruction alignment across all modalities and base models, particularly enhancing content fidelity, as further visualized in Fig. 2(c).

Model-centric analysis. Results consistently confirm AnyCap’s instruction alignment gains across both the 2B and 8B variants. Gains also extend to large-capacity InternVL3.5-38B. Notably, with AnyCap, certain open-source models achieve instruction alignment scores that are comparable to or even exceed strong unassisted proprietary base-lines. For example, AnyCap-2B elevates the performance of open-source InternVL2.5 from ~ 2.1 to ~ 2.4 in the image style category. The enhanced score surpasses the ~ 2.3 achieved by GPT-4o. Extended results with additional base models are in Appendix C.

Task-centric analysis. Boosts in content-related instruction alignment typically surpass those in style, likely because content controls provide more explicit learning signals, whereas stylistic constraints are inherently more subjective and harder to en-

Table 2: Image captioning on AnyCapEval. Abbreviations like IPos. are varying instruction types in Table 1. Content and style values are computed using KPD and style scores. Gains are highlighted in blue font.

Model	Content ↑					Style ↑					
	IPos.	IApp.	Ins.	Per.	Avg.	Brf.	Det.	Thm.	Poe.	Nar.	Avg.
<i>Proprietary Models</i>											
GPT-4o	1.55	3.25	3.84	2.94	2.89	1.91	1.86	2.59	2.74	2.59	2.26
AnyCap-2B	3.01 _{1.46}	4.32 _{1.07}	4.55 _{0.71}	4.57 _{1.62}	4.11 _{1.22}	2.29 _{0.38}	1.98 _{0.12}	2.38 _{0.21}	3.02 _{0.29}	2.78 _{0.20}	2.46 _{0.19}
AnyCap-8B	3.48 _{1.93}	4.81 _{1.56}	4.89 _{1.06}	5.00 _{2.05}	4.54 _{1.65}	2.38 _{0.47}	2.35 _{0.49}	2.82 _{0.24}	3.10 _{0.37}	2.87 _{0.28}	2.65 _{0.39}
<i>Open-source Models</i>											
InternVL2.5-8B	1.51	3.43	4.54	2.68	3.04	2.13	1.92	1.94	2.08	2.52	2.12
AnyCap-2B	3.22 _{1.71}	4.49 _{1.06}	4.82 _{0.28}	4.00 _{1.32}	4.13 _{1.09}	2.27 _{0.14}	1.78 _{0.14}	2.38 _{0.44}	2.71 _{0.63}	2.67 _{0.15}	2.36 _{0.24}
Qwen2.5VL-7B	1.51	3.52	5.18	3.03	3.31	2.00	2.02	2.50	2.14	2.63	2.22
AnyCap-2B	3.38 _{1.87}	4.42 _{0.90}	4.94 _{0.24}	4.68 _{1.65}	4.36 _{1.05}	2.40 _{0.40}	2.04 _{0.02}	2.50 _{0.00}	2.94 _{0.80}	2.72 _{0.09}	2.50 _{0.28}
InternVL3.5-8B	1.81	3.90	4.98	3.07	3.49	2.10	1.94	2.23	2.37	2.54	2.22
AnyCap-2B	3.26 _{1.45}	4.50 _{0.60}	5.56 _{0.58}	4.24 _{1.16}	4.39 _{0.90}	2.10 _{0.00}	1.94 _{0.00}	2.38 _{0.16}	2.88 _{0.51}	2.61 _{0.07}	2.33 _{0.10}
InternVL3.5-14B	1.96	3.91	4.59	3.48	3.49	2.18	2.08	2.24	2.57	2.59	2.31
AnyCap-2B	3.17 _{1.21}	4.66 _{0.75}	5.57 _{0.98}	3.85 _{0.37}	4.31 _{0.82}	2.18 _{0.00}	1.88 _{0.20}	2.35 _{0.12}	2.90 _{0.33}	2.63 _{0.04}	2.33 _{0.03}
InternVL3.5-38B	1.96	4.15	4.50	3.65	3.56	2.16	2.16	2.24	2.51	2.59	2.30
AnyCap-2B	3.16 _{1.20}	4.39 _{0.24}	4.65 _{0.15}	4.07 _{0.42}	4.07 _{0.51}	2.18 _{0.02}	1.98 _{0.18}	2.24 _{0.00}	2.86 _{0.35}	2.74 _{0.15}	2.36 _{0.06}

Table 3: Video captioning on AnyCapEval. Appendix C extends our evaluation with more base models and detailed results, validating the effectiveness of AnyCap.

Model	Content ↑					Style ↑					
	IPos.	IApp.	IAct.	Mov.	Evt.	Avg.	Brf.	Det.	Poe.	Nar.	Avg.
<i>Proprietary Models</i>											
GPT-4o	2.41	4.00	3.86	2.70	3.03	3.55	1.47	1.52	2.52	2.48	2.15
AnyCap-2B	3.45 _{1.04}	6.15 _{2.15}	6.26 _{2.40}	6.53 _{3.83}	4.60 _{1.57}	5.30 _{1.75}	1.97 _{0.50}	1.78 _{0.26}	2.60 _{0.08}	2.65 _{0.17}	2.30 _{0.15}
AnyCap-8B	4.92 _{2.51}	6.60 _{2.60}	7.68 _{3.82}	5.67 _{2.97}	4.81 _{1.78}	5.74 _{2.19}	1.95 _{0.48}	1.82 _{0.30}	2.50 _{0.02}	2.77 _{0.29}	2.32 _{0.17}
<i>Open-source Models</i>											
InternVL2.5-8B	3.08	4.33	3.44	3.16	2.55	3.52	1.39	1.77	1.91	2.34	1.93
AnyCap-2B	4.67 _{1.59}	6.32 _{1.99}	3.19 _{0.25}	6.49 _{3.33}	4.50 _{1.95}	5.07 _{1.55}	1.84 _{0.45}	1.80 _{0.03}	2.48 _{0.57}	2.39 _{0.05}	2.19 _{0.26}
Qwen2.5VL-7B	2.88	4.20	3.29	2.24	2.93	3.54	1.63	1.65	2.40	2.29	2.10
AnyCap-2B	3.39 _{0.51}	5.97 _{1.77}	6.23 _{2.94}	6.30 _{4.06}	4.48 _{1.55}	5.25 _{1.71}	1.66 _{0.03}	1.82 _{0.17}	2.30 _{0.10}	2.55 _{0.26}	2.16 _{0.06}
InternVL3.5-8B	3.14	4.53	5.05	2.93	3.17	3.76	1.43	1.80	2.12	2.41	1.94
AnyCap-2B	4.20 _{1.06}	5.88 _{1.35}	6.40 _{1.35}	5.58 _{2.65}	4.72 _{1.55}	5.36 _{1.60}	1.82 _{0.40}	1.86 _{0.06}	2.31 _{0.19}	2.44 _{0.03}	2.11 _{0.17}
InternVL3.5-14B	3.23	4.63	5.16	3.03	3.18	3.85	1.45	1.86	2.21	2.48	2.00
AnyCap-2B	3.37 _{0.14}	5.44 _{0.81}	6.36 _{1.20}	4.83 _{1.80}	4.42 _{1.24}	4.88 _{1.03}	1.81 _{0.36}	2.16 _{0.30}	2.33 _{0.12}	2.54 _{0.06}	2.21 _{0.21}
InternVL3.5-38B	3.31	4.77	5.01	4.53	3.40	4.20	1.45	2.05	2.28	2.52	2.08
AnyCap-2B	3.59 _{0.28}	5.52 _{0.75}	6.04 _{1.03}	4.84 _{0.31}	4.18 _{0.78}	4.83 _{0.63}	1.93 _{0.48}	2.20 _{0.15}	2.55 _{0.27}	2.57 _{0.05}	2.31 _{0.23}

Table 4: Audio captioning on AnyCapEval.

Model	Content ↑		Style ↑			
	Evt.	Brf.	Nar.	Poe.	Avg.	
<i>Proprietary Models</i>						
GPT-4o	1.59	1.42	1.24	0.88	1.18	
AnyCap-2B	1.79 _{0.20}	1.44 _{0.02}	1.38 _{0.14}	1.00 _{0.12}	1.28 _{0.10}	
AnyCap-8B	1.88 _{0.29}	1.42 _{0.00}	1.40 _{0.16}	1.08 _{0.20}	1.30 _{0.12}	
GPT-4o mini	1.28	1.17	1.16	0.79	1.04	
AnyCap-2B	1.56 _{0.28}	1.21 _{0.04}	1.25 _{0.09}	0.89 _{0.10}	1.11 _{0.07}	
<i>Open-source Models</i>						
MiniCPM-o	1.42	1.02	0.68	0.43	0.71	
AnyCap-2B	1.55 _{0.13}	1.37 _{0.35}	1.32 _{0.64}	0.92 _{0.49}	1.21 _{0.50}	

Table 5: Performance on MIA-Bench.

Model	desc.	ment.	len.	persp.	gen.	avg.
GPT-4o	90.3	87.9	90.5	85.2	91.6	89.1
+AnyCap-2B	89.1 _{1.2}	88.9 _{1.4}	90.3 _{1.2}	90.0 _{1.8}	92.2 _{0.6}	90.1 _{1.0}
+AnyCap-8B	90.4 _{0.1}	89.4 _{1.5}	89.9 _{0.6}	89.2 _{1.0}	92.9 _{1.3}	90.3 _{1.2}
InternVL2.5-8B	85.2	79.0	82.6	72.5	85.2	80.9
+AnyCap-2B	85.7 _{0.5}	79.6 _{0.6}	85.2 _{0.6}	77.5 _{5.0}	85.7 _{0.5}	82.7 _{1.8}
Yi-VL-34B	71.7	52.0	55.8	60.0	59.8	59.9
+AnyCap-2B	71.8 _{0.1}	54.2 _{2.2}	61.9 _{6.1}	61.7 _{1.7}	61.0 _{1.2}	62.1 _{2.2}

Table 6: Performance on VidCapBench.

Model	acc. ↑	pre. ↑	cov. ↑	con. ↑
GPT-4o	15.6	58.9	87.8	6.3
+AnyCap-2B	15.4 _{0.2}	59.1 _{0.2}	87.1 _{0.7}	10.3 _{4.0}
+AnyCap-8B	15.8 _{0.2}	59.5 _{0.6}	88.0 _{0.2}	9.3 _{3.0}
InternVL2.5-8B	12.8	51.6	84.2	7.1
+AnyCap-2B	13.8 _{1.0}	55.7 _{4.1}	84.7 _{0.5}	10.1 _{1.0}
Qwen2.5VL-7B	13.1	53.5	82.9	6.6
+AnyCap-2B	14.6 _{1.5}	55.8 _{2.3}	85.0 _{2.1}	10.6 _{4.0}

Table 7: Human evaluation on AnyCapEval comparing manually written and AnyCap.

Method	Image ↑		Video ↑		Audio ↑	
	Cont.	Sty.	Cont.	Sty.	Cont.	Sty.
Expert	4.62	2.77	6.04	2.62	3.30	4.13
Non-expert	4.42	2.66	5.72	2.35	2.01	1.37
GPT+AnyCap	4.54	2.65	5.74	2.32	1.88	1.30

force. Visual tasks outperform audio counterparts, potentially due to the richer supervision and higher information density in vision-language pretraining. AnyCap exhibits varied effects across settings: in well-studied tasks (e.g., image appearance), it refines already strong captions; in less-explored, complex tasks (e.g., video actions or audio events), it fills in missing content, yielding larger gains.

4.2 Evaluation on Public Benchmarks

To assess the cross-modal generalizability of AnyCap, we integrate it with diverse backbones and evaluate its performance on a series of public image, video, and audio captioning benchmarks. On MIA-Bench (Qian et al., 2025) (Table 5), AnyCap consistently improves the instruction-following alignment across models such as InternVL2.5 and Yi-VL (Young et al., 2024). Furthermore, integrating AnyCap-8B with GPT-4o yields additional gains, achieving the best performance under our

evaluation setting. On VidCapBench (Chen et al., 2025b) (Table 6), augmenting Qwen2.5VL-7B with AnyCap yields improvements in accuracy (+1.5), precision (+2.3), and conciseness (+4.0), indicating fewer hallucinations and better capture of motion and content cues. For the audio modality, although the ground-truth references are shorter and metrics more sensitive on Clotho and AudioCaps (Drossos et al., 2020; Kim et al., 2019) (Table 8), AnyCap still consistently enhances fluency and semantic controllability. This also reflects that instruction-aligned audio captioning remains comparatively underdeveloped relative to image/video: open-source audio baselines are markedly weaker (e.g., MiniCPM-o’s AnyCapEval-Audio style score averages only 0.71, versus ≥ 1.04 for proprietary models in Table 4), so AnyCap’s relative gains in this modality are correspondingly larger (up to +70% relative improvement in style for MiniCPM-o), underscoring AnyCap’s contribution toward ad-

Table 8: Performance on AudioCaps and Clotho. Metrics: spice (sp), sentence-bert (sbt).

Model	AudioCaps $\uparrow$		Clotho $\uparrow$	
	sp.	sbt.	sp.	sbt.
GPT-4o	0.06	0.41	0.07	0.37
AnyCap-2B	0.06 _{0.00}	0.45 _{0.04}	0.07 _{0.00}	0.40 _{0.03}
AnyCap-8B	0.07 _{0.01}	0.45 _{0.04}	0.08 _{0.01}	0.40 _{0.03}

Table 10: AnyCap vs. SFT, DPO, SC on MIA benchmark.

SFT	DPO	SC	AnyCap
1.2%	+7.0%	+2.8%	+16.9%

Table 11: Comparison of the effect of different data ratios (q , uncontrolled a , c), (q , c , c), and (q , hallucinated a , c).

1:1:1	1:2:2	2:1:2	2:2:1
+17.9%	+17.2%	+16.5%	+21.1%

vancing this comparatively underexplored modality.

4.3 Component-wise Analysis

Comparison with SFT, DPO, and Self-Critic approaches. One motivation for plug-and-play AnyCap is to avoid retraining each base model, as required by fine-tuning and RL-based policy optimization approaches (Luo et al., 2023; Xu et al., 2026). We compare it against three options: (i) supervised fine-tuning (SFT), where instructions are added to prompts to fine-tune the base model; (ii) direct preference optimization (DPO) on base models using preferred–non-preferred pairs from AnyCapData; and (iii) a self-critic (SC) approach where the backbone revises its own outputs at test time. Across both AnyCapEval (Table 9) and MIA (Table 10), AnyCap consistently achieves stronger performance. This advantage arises as it builds on base captions as references, focusing learning on refining outputs toward user instructions rather than generating fully aligned captions from scratch. Detailed baseline settings are in Appendix C.1. Table 12 further confirms that AnyCap performs targeted revision (pruning irrelevant content while preserving keypoints) rather than full unconditional re-captioning of the input.

Ablation on training data ratio. Our triplet training data consist of three main types: (q , uncontrolled a , c), (q , c , c), and (q , hallucinated a , c). As shown in Fig. 5, performance peaks at $\sim 40\%$ fully correct samples then declines when exceeding 50%, suggesting that excessive reliance may cause overfitting. We further perform ablation across all

Table 9: AnyCap vs. SFT, DPO, SC on AnyCapEval. Entries are relative improvements (%) over each model’s origin baseline under different paradigms.

Model (Modality)	SFT		DPO		SC		AnyCap-2B		AnyCap-8B	
	Cont.	Sty.	Cont.	Sty.	Cont.	Sty.	Cont.	Sty.	Cont.	Sty.
InternVL2.5-8B(Ima)	+9.2	+3.5	+9.5	+4.4	+3.3	+0.5	+35.9	+11.3	+40.1	+16.0
InternVL2.5-8B(Vid)	+2.3	+1.0	+2.8	+1.6	+0.0	-1.6	+44.0	+13.5	+62.8	+16.1
Qwen2.5VL-7B(Ima)	+6.7	+3.8	+7.5	+4.1	+11.5	+5.0	+31.7	+12.6	+36.9	+15.3
Qwen2.5VL-7B(Vid)	+3.2	+3.0	+7.3	+5.2	+16.7	+11.0	+26.6	+2.9	+56.8	+10.0



Figure 4: Human evaluation comparing AnyCap-8B with GPT-4o. Captions refined with AnyCap align more closely with the given instructions.

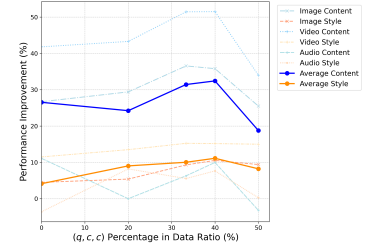


Figure 5: Impact of (q , c , c) percentage in training data ratio.

Table 12: Absolute keypoint counts (no length normalization) and response lengths on image captioning.

Model	Keypoints (Abs.) $\uparrow$				Response Length			
	Per.	Ins.	IApp.	IPos.	Per.	Ins.	IApp.	IPos.
GPT-4o	2.96	3.52	3.13	1.42	95.8	91.8	92.4	93.7
+AnyCap	2.29	4.00	3.13	1.55	49.6	83.3	61.7	51.8
InternVL2.5-8B	2.54	3.64	2.85	1.21	90.0	84.6	83.0	68.0
+AnyCap	2.44	3.76	2.85	1.70	62.5	76.6	59.4	49.8
Qwen2.5VL-7B	2.68	3.80	2.82	1.21	84.0	70.3	79.3	74.2
+AnyCap	2.59	3.78	2.95	1.64	64.0	68.0	66.7	51.6

three types. Results in Table 11 show that including more uncontrolled a samples yields better performance than emphasizing hallucinated a . A likely reason is that outputs from strong base models are more prone to instruction violations than factual errors, so AnyCap gains more from the process of frequently correcting uncontrolled cases.

Keypoint coverage vs. Caption length. Table 12 shows AnyCap does not “win by writing more” but effectively reduces unrelated details (especially from GPT-4o, as in Fig. 6) while maintaining or improving keypoint accuracy, explaining its consistently higher AnyCapEval scores across all instruction types and modalities.

4.4 Human Evaluation

To complement model-based evaluation, we conduct two-fold human study. First, a blind preference test (Fig. 4) compares AnyCap-8B with GPT-4o on content and style criteria. Results confirm our refinement bridges model outputs and human intentions. Second, we hire PhD-level experts and well-educated non-experts to handwrite captions for comparison with AnyCap. As shown in Table 7, AnyCap matched or slightly outperformed the non-

expert group on visual tasks and remained broadly comparable on audio, suggesting that it delivers strong quality while aligning well with human preferences. Evaluation reliability analyses (judge replacement, prompt robustness) are detailed in Appendix D.

5 Conclusion

We present AnyCap, a lightweight plug-and-play module that aligns base captioners with human instructions across images, videos, and audio without retraining base models. A single model serves all three modalities and applies to both proprietary and open-source backbones. We also introduce AnyCapData, a large-scale instruction-based caption dataset, and AnyCapEval, a benchmark for fine-grained instruction-alignment evaluation. AnyCap consistently improves caption quality and instruction adherence on both AnyCapEval and public benchmarks across all modalities.

Limitations. While AnyCap demonstrates strong performance across image, video, and audio modalities, its primary limitation lies in the availability of high-quality data resources for emerging modalities, such as 3D structures or tactile signals. However, since our framework is inherently modular and extensible, it can be readily adapted to these domains once the corresponding datasets are available. Given sufficient data resources, researchers can easily scale AnyCap to broader omni-modal scenarios by following our open-sourced implementation and training protocols.

Ethics statement. The data collection and construction of AnyCapData strictly adhere to the original licenses of the source datasets (e.g., CC BY 4.0). We implemented a manual filtering process to ensure the removal of all content unsuitable for the workplace, offensive content, or personally identifiable information. Despite our best efforts, subtle biases related to cultural or regional differences may still exist due to the inherent distribution of the underlying web-scale data. Therefore, we encourage the community to use our resources responsibly and abide by our open-source licensing agreement. To promote transparency and collective oversight, we call upon all researchers conducting research based on our work to release their code and data to ensure that the development of controllable image captioning technology is ethically sound and benefits the broader society.

Acknowledgements. This work was partly sup-

ported by the National Key Research and Development Program of China (No. 2024YFB2808903), the Shenzhen Science and Technology Program (ZDCY20250901103533010). This work is supported by Shanghai Artificial Intelligence Laboratory.

References

- Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, and 33 others. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.
- Luigi Celona, Simone Bianco, Marco Donzella, and Paolo Napoletano. 2023. Improving image captioning descriptiveness by ranking and llm-based fusion. *CoRR*, abs/2306.11593.
- Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. 2024. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*.
- David M. Chan, Suzanne Petryk, Joseph E. Gonzalez, Trevor Darrell, and John F. Canny. 2023. CLAIR: evaluating image captions with large language models. In *EMNLP*, pages 13638–13646. Association for Computational Linguistics.
- David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200. The Association for Computer Linguistics.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin.

- 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV (17)*, volume 15075 of *Lecture Notes in Computer Science*, pages 370–387. Springer.
- Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. 2024b. EAT: self-supervised pre-training with efficient audio transformer. In *IJCAI*, pages 3807–3815. ijcai.org.
- Xiaofu Chen, Israfel Salazar, and Yova Kementchedjhiya. 2025a. Specs: Specificity-enhanced clip-score for long image caption evaluation. *arXiv preprint arXiv:2509.03897*.
- Xinlong Chen, Yuanxing Zhang, Chongling Rao, Yushuo Guan, Jiaheng Liu, Fuzheng Zhang, Chengru Song, Qiang Liu, Di Zhang, and Tieniu Tan. 2025b. Vidcapbench: A comprehensive benchmark of video captioning for controllable text-to-video generation. *CoRR*, abs/2502.12782.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*, pages 8307–8316. Computer Vision Foundation / IEEE.
- Erfei Cui, Yinan He, Zheng Ma, Zhe Chen, Hao Tian, Weiyun Wang, Kunchang Li, Yi Wang, Wenhui Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yali Wang, Limin Wang, Yu Qiao, and Jifeng Dai. 2024. [Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o](#).
- Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. Benchmarking and improving detail image caption. *CoRR*, abs/2405.19092.
- Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *ICASSP*, pages 736–740. IEEE.
- Debidatta Dwibedi, Vidhi Jain, Jonathan Tompson, Andrew Zisserman, and Yusuf Aytar. 2024. Flexcap: Describe anything in images in controllable detail. In *NeurIPS*.
- Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Zhenheng Yang, Chaoyou Fu, Xiang Li, Jian Yang, and Ying Tai. 2024. Instancecap: Improving text-to-video generation via instance-aware structured caption. *CoRR*, abs/2412.09283.
- Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. 2024. Visual fact checker: Enabling high-fidelity detailed caption generation. In *CVPR*, pages 14033–14042. IEEE.
- Wei Han, Hui Chen, and Soujanya Poria. 2024. Towards robust instruction tuning on multimodal large language models. *arXiv preprint arXiv:2402.14492*.
- Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X. Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on LLM performance? *CoRR*, abs/2411.10541.
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. In *NeurIPS*.
- Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan Wang, and Zicheng Liu. 2024. Segment and caption anything. In *CVPR*, pages 13405–13417. IEEE.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.
- Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Qiu, Juntao Dai, and Yaodong Yang. 2024. Aligner: Efficient alignment by learning to correct. In *NeurIPS*.
- Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaying Zhang, Tetsuya Sakai, and Yujiu Yang. 2023. MAP: multimodal uncertainty-aware vision-language pre-training model. In *CVPR*, pages 23262–23271. IEEE.
- Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024. Miradata: A large-scale video dataset with long durations and structured captions. In *NeurIPS*.
- Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT (1)*, pages 119–132. Association for Computational Linguistics.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuoqun Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duoqun Huang, Fang Yang, and 33 others. 2024. Hunyuanvideo: A systematic framework for large video generative models. *CoRR*, abs/2412.03603.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017a. Dense-captioning events in videos. In *ICCV*, pages 706–715. IEEE Computer Society.

- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017b. Dense-captioning events in videos. In *ICCV*, pages 706–715. IEEE Computer Society.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP*, pages 2511–2522. Association for Computational Linguistics.
- Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. 2024. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. *CoRR*, abs/2412.08614.
- Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *CoRR*, abs/2308.08747.
- Koki Maeda, Shuhei Kurita, Taiki Miyanishi, and Naoaki Okazaki. 2024. Vision language model-based caption evaluation method leveraging visual context extraction. *CoRR*, abs/2402.17969.
- Shunqi Mao, Chaoyi Zhang, Hang Su, Hwanjun Song, Igor Shalyminov, and Weidong Cai. 2024. Controllable contextualized image captioning: Directing the visual narrative through user-defined highlights. In *ECCV (50)*, volume 15108 of *Lecture Notes in Computer Science*, pages 464–481. Springer.
- Xinhao Mei, Xubo Liu, Mark D. Plumbley, and Wenwu Wang. 2022. Automated audio captioning: an overview of recent progress and new challenges. *EURASIP J. Audio Speech Music. Process.*, 2022(1):26.
- Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. 2024. DOCCI: descriptions of connected and contrasting images. In *ECCV (60)*, volume 15118 of *Lecture Notes in Computer Science*, pages 291–309. Springer.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.
- Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. 2025. Mia-bench: Towards better instruction following evaluation of multimodal llms. In *ICLR*. OpenReview.net.
- Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. 2023. Gender biases in automatic evaluation metrics for image captioning. In *EMNLP*, pages 8358–8375. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125.
- Yiming Ren, Junjie Wang, Yuxin Meng, Yihang Shi, Zhiqiang Lin, Ruihang Chu, Yiran Xu, Ziming Li, Yunfei Zhao, Zihan Wang, Yu Qiao, Ruiming Tang, Minghao Liu, and Yujiu Yang. 2026. SIN-Bench: Tracing native evidence chains in long-context multimodal scientific interleaved literature. *CoRR*, abs/2601.10108.
- Dhruv Sharma, Chhavi Dhiman, and Dinesh Kumar. 2023. Evolution of visual data captioning methods, datasets, and evaluation metrics: A comprehensive survey. *Expert Syst. Appl.*, 221:119773.
- Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, and 1 others. 2025. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *ACL (Findings)*, pages 13088–13110. Association for Computational Linguistics.
- Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2024. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *CVPR*, pages 26690–26699. IEEE.

- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575. IEEE Computer Society.
- Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence - video to text. In *ICCV*, pages 4534–4542. IEEE Computer Society.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164. IEEE Computer Society.
- Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, and 43 others. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.
- Dingdong Wang, Jin Xu, Ruihang Chu, Zhifang Guo, Xiong Wang, Jincenzi Wu, Dongchao Yang, Shengpeng Ji, and Junyang Lin. 2025a. Inserter: Speech instruction following with unsupervised interleaved pre-training. In *ACL*, pages 18024–18046. Association for Computational Linguistics.
- Jiawei Wang, Liping Yuan, Yuchen Zhang, and Hao-miao Sun. 2024a. Tarsier: Recipes for training and evaluating large video description models. *CoRR*, abs/2407.00634.
- Teng Wang, Jinrui Zhang, Junjie Fei, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, and Shanshan Zhao. 2023. Caption anything: Interactive image description with diverse multimodal controls. *CoRR*, abs/2305.02677.
- Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, Yu Qiao, and Jifeng Dai. 2024b. The all-seeing project V2: towards general relation comprehension of the open world. In *ECCV (33)*, volume 15091 of *Lecture Notes in Computer Science*, pages 471–490. Springer.
- Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, and 1 others. 2024c. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *ICLR*.
- Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, and 1 others. 2020. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3268–3278.
- Zhen Wang, Jun Xiao, Yueting Zhuang, Fei Gao, Jian Shao, and Long Chen. 2025b. Learning combinatorial prompts for universal controllable image captioning. *Int. J. Comput. Vis.*, 133(1):129–150.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296. IEEE Computer Society.
- Yiran Xu, Yiming Ren, Zicheng Lin, Chufan Shi, Yukang Chen, Dingdong Wang, Tianhe Wu, Junjie Wang, Yujiu Yang, Yu Qiao, and Ruihang Chu. 2026. Smaller models are natural explorers for policy-level diversity in GRPO. *CoRR*, abs/2605.30789.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A GPT-4V level MLLM on your phone. *CoRR*, abs/2408.01800.
- Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, and 11 others. 2024. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78.
- Zejun Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. 2023. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *CVPR*, pages 23465–23476. IEEE.
- Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *CoRR*, abs/2503.05379.
- Yuzhong Zhao, Yue Liu, Zonghao Guo, Weijia Wu, Chen Gong, Qixiang Ye, and Fang Wan. 2024. Controlcap: Controllable region-level captioning. In *ECCV (38)*, volume 15096 of *Lecture Notes in Computer Science*, pages 21–38. Springer.

Appendix

A Datasets

A.1 Overview of Caption Datasets

Automatic captioning for modalities such as images, videos, and audio is a core task in artificial intelligence, aiming to enable machines to understand and describe perceived content using natural language, much like humans do. The foundation supporting this research is large-scale annotated datasets. However, existing datasets (Table 13) exhibit significant limitations in providing fine-grained, multi-dimensional control capabilities. We next detail the motivation and data analysis behind our dataset.

A.2 The Singularity and Scarcity of User Instructions

A “user instruction” is an input accompanying the main media (image, video, or audio), intended to guide or constrain the generation of descriptive content. Ideally, we desire models capable of generating descriptions with varying styles, focusing on different aspects, and meeting specific requirements based on diverse user instructions.

Most classic datasets, such as COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014) in the image domain, MSR-VTT (Xu et al., 2016) and MSVD (Chen and Dolan, 2011) in the video domain, and AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) in the audio domain, primarily provide “media-caption pairs”. They typically include multiple reference captions provided by different annotators. This “multi-reference” design is mainly intended to evaluate the diversity and coverage of descriptions, rather than providing explicit control during generation. The model’s primary task is to generate a “reasonable” description, but the user cannot specify requirements at generation time (e.g., “generate a humorous caption”, “describe the background elements in detail”). The only implicit user instruction is essentially “describe this content in brief”.

Some datasets offer a degree of control, but often only along a single dimension. Dense captioning datasets, like DCI (Urbanek et al., 2024) for images or video datasets providing temporal information (Krishna et al., 2017b), associate descriptions with specific image regions or video segments. The “region/time segment” here can be viewed as a user instruction, i.e., “describe *this part*”. However, this

is usually the sole dimension of control. Datasets like ASD v2 (Wang et al., 2024b) focus on describing relationships between entities in an image. The user instruction is the “entity pair” whose relationship needs to be described. Datasets guided by concepts like MiraData (Ju et al., 2024) or InstanceCap (Fan et al., 2024) require the output to follow a specific structure, which constitutes a form of control over the output format. Although such structured outputs often encompass several instruction types (e.g., Background, Detail), the formulation and granularity of these user instructions are typically fixed, lacking the flexibility for customized control requests tailored to each individual image, video or audio.

In summary, the majority of existing public datasets lack the ability to align captions with specific user instructions across multiple dimensions, limiting the development and evaluation of instruction-aligned captioning technologies. Even when leveraging large language models such as GPT for annotation, the resulting captions often reflect general patterns and fail to align precisely with the specified instructions along desired dimensions.

To address this gap, we construct AnyCapData, a large-scale dataset featuring diverse user instructions paired with high-quality, instruction-compliant captions across multiple modalities. Leveraging this dataset, we train AnyCapModel to achieve fine-grained instruction-aligned caption generation. As illustrated in Fig. 6, AnyCap can transform initial captions into highly specific descriptions that accurately follow various user instructions, enabling precise and customizable outputs without requiring modifications to the base model architecture.

A.3 Data Analysis

Data sources. To construct AnyCapData, we utilize multiple publicly available datasets spanning three modalities: images, videos, and audio. Specifically, we incorporate data from ASD v2, DCI, DOCCI, ShareGPT-4o, InstanceCap, and AudioCaps. From these datasets, we extract images, videos, and audio content, and employ various multimodal large language models (MLLMs) to generate diverse user instructions and captions based on predefined instructions.

Data statistics. AnyCapData encompasses three primary modalities: images, videos, and audio, comprising a total of 300k (q, c, a) triplets. Among

Table 13: Overview of the publicly available caption datasets analyzed in this study. “Modality” indicates the primary input type (e.g., image, video, or audio), while “Instruction Type” refers to the predominant linguistic or structural annotation style as described by the original dataset authors.

Dataset	Modality	Instruction Type
ASD v2	Image	Relation
MDVP-Data	Image	Region / Brief / Detail
DCI	Image	Dense
DOCCI	Image	Dense
ImageInWords (IIW)	Image	Dense
ShareGPT-4o	Image	Detail
ShareGPT-4v	Image	Detail
ShareGPT-4o (video)	Video	Detail
MSR-VTT	Video	Brief
MSVD (YouTube2Text)	Video	Brief
VATEX	Video	Brief
Video-ChatGPT (VideoInstruct-100K)	Video	Dense
InstanceCap / InstanceVid	Video	Instance-level (structured)
ShareGPT4Video	Video	Detail
MiraData	Video	Structured
LLaVA-Video-178K	Video	Detail
AudioCaps	Audio	Brief
Clotho	Audio	Brief
MACS	Audio	Brief
WavCaps	Audio	Brief

them, 125k belong to the image modality, 100k to video modality, and 75k to audio modality. Each modality includes multiple instruction types.

Generation of instruction-caption triplets. To generate the (q, c, a) triplets in AnyCapData, where q denotes a user instruction, c is a high-quality caption, and a is a relatively suboptimal caption, we utilize several MLLMs guided by specifically designed instruction templates. The detailed templates employed in this work are presented in Table 14 and Table 15. These templates are constructed according to different instruction types to ensure structured generation and facilitate downstream analysis. Each raw data sample is paired with one or more tailored instructions to produce diverse outputs.

A.4 Future Directions

Future dataset development could focus on the following improvements:

Constructing datasets with rich user instructions: There is a need to design new datasets containing diverse, explicit, and composable instruction types.

Leveraging synthetic data: Explore the use of

large language models to generate synthetic description data with specific user instructions, serving as a supplement to real data.

Improving evaluation methods: Develop new evaluation protocols and metrics capable of assessing fine-grained aspects like controllability, style consistency, and faithfulness.

B AnyCapEval Benchmark

B.1 Limitations of Existing Evaluation Methods

Limitations of traditional machine translation metrics. Metrics developed for machine translation, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015), primarily focus on n-gram overlap or co-occurrence statistics. While they can measure fluency and lexical similarity, they fail to accurately capture semantic consistency. For instance, sentences like “I do love large models” and “I do not love large models” could receive similarly high scores, despite conveying opposite meanings. These metrics cannot effectively detect deep semantic divergence, factual errors, or noncompliance with user instructions.

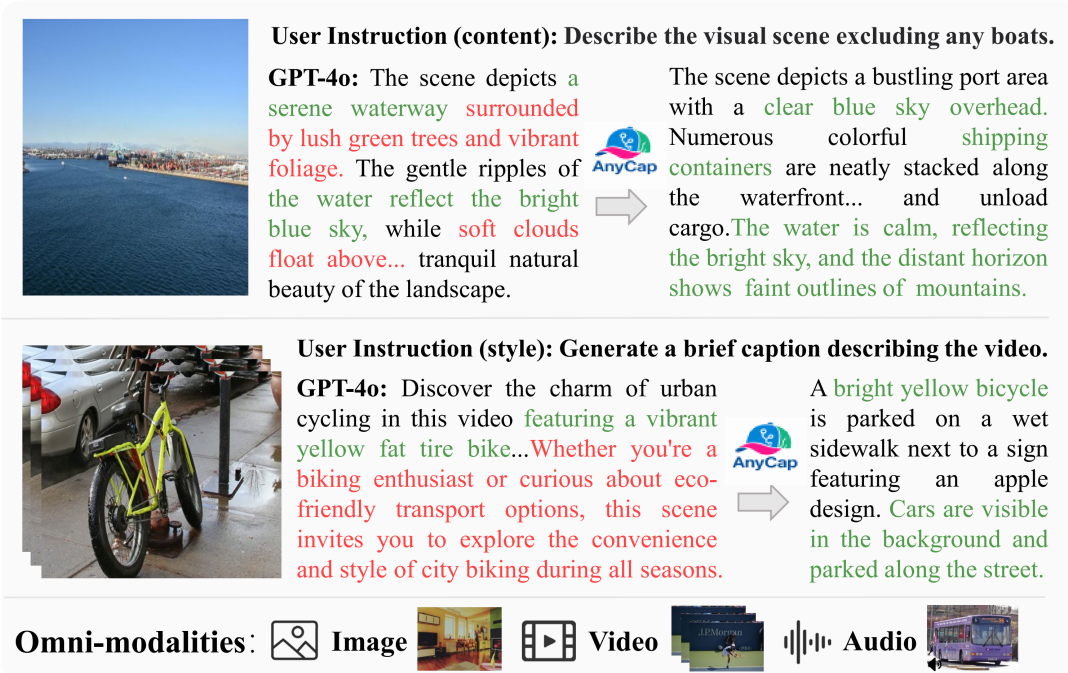


Figure 6: AnyCap enables controllable captioning across modalities by refining base model outputs to better align with user instructions. Given a user instruction, it takes initial captions from a foundation model and corrects instruction violations (highlighted in red), producing compliant, instruction-following outputs (green), all without requiring fine-tuning of the base model.

Challenges of multimodal large language models scoring. Recent evaluation schemes leverage powerful language models or multimodal large language models (MLLMs) to directly score captions. However, these methods suffer from significant randomness and instability: the same caption might receive drastically different scores across minor prompt variations or multiple queries. Additionally, hallucinated or semantically incorrect captions may still receive high scores, undermining the reliability of such automatic evaluation.

Instability of keypoint extraction-based metrics. Some studies, including those utilizing benchmarks like (Wang et al., 2024a) which aims for fine-grained descriptions, propose extracting key information points from captions using MLLMs and computing precision, recall, and F1 scores. However, keypoint extraction itself is unstable: the number and granularity of extracted points vary widely across runs. Our empirical analysis shows low correlation between model-extracted keypoints and human-annotated ones, suggesting that such extraction-based metrics poorly reflect the true information content of captions.

Challenges of QA-pair based evaluation. Benchmarks such as VDC (Chai et al., 2024) highlight that QA-pair based evaluation suffers from instability, primarily due to the variable quality and granularity of the QA pairs themselves. Designing high-quality, representative QA items that faithfully capture fine-grained caption attributes remains an open challenge.

B.2 Rationale Behind AnyCapEval Design

B.2.1 Separation of Content and Style

Through extensive analysis of user instruction types in captioning, we divide evaluation into two orthogonal dimensions:

- **Content:** We expect the generated captions to strictly adhere to the given user instructions, maintain clear topical relevance, and avoid introducing irrelevant information. To evaluate this aspect, we employ keypoint density (KPD) as the primary metric.
- **Style:** We aim for the generated captions to closely match human-crafted references in terms of narrative style, tone, and expressive form. To assess this, we design a fine-grained scoring system.

This separation is crucial because content and style represent fundamentally different aspects of caption quality that require distinct evaluation approaches. Content adherence focuses on factual accuracy and instruction compliance: whether the caption correctly follows the user’s specified requirements about what to describe. This is objectively verifiable through user instruction alignment. In contrast, style evaluation assesses subjective qualities like fluency, coherence, and naturalness that reflect human-like expression. Merging these dimensions would conflate objective and subjective criteria, making it difficult to diagnose specific model weaknesses. The orthogonal evaluation allows for more precise identification of whether a model’s limitations lie in its ability to follow instructions (content) or in its linguistic quality (style), enabling targeted improvements to each aspect. Furthermore, different applications may prioritize these dimensions differently: some use cases demand strict content control while others emphasize stylistic quality, making separate evaluation essential for practical deployment decisions.

B.2.2 Keypoint Density Metric

Intuitive motivation. We expect the generated captions to efficiently embed as many control-required key information points as possible within a limited textual length, while minimizing irrelevant or redundant content. This analogy is comparable to adding an appropriate amount of solute into a limited solvent to achieve a targeted concentration. Excessively diluting it in a large solvent volume, even if the absolute quantity increases, results in negligible concentration and ineffective outcomes. To prevent models from exploiting this metric by generating excessively short captions that artificially inflate density scores, we normalize the number of extracted keypoints by the caption length (or an appropriate normalization factor) and apply penalties to captions whose lengths fall outside the predefined acceptable range.

Length Normalization. We randomly sample approximately 200 caption instances and extract information points using both GPT-4o (including both control-required keypoints and irrelevant information points) and human annotators. The human-annotated information points are obtained by post-processing the model-generated information points through manual filtering and revision. Correlation analyses, presented in the main paper, reveal that

caption length exhibits a significantly stronger association with human-annotated information content than the number of keypoints automatically extracted by the model. This empirical evidence further supports the rationality of applying length normalization in the evaluation metric.

B.2.3 GPT-4o As The Evaluator

When using GPT-4o as the evaluator, we carefully design structured prompts that explicitly instruct the model to assess Content and Style separately, ensuring consistent and interpretable judgments. See Table 17 and Table 18 for the detailed prompts used for Content and Style evaluation.

C Additional Experimental Details

C.1 Experimental Setup

The experimental setup utilizes a computing infrastructure comprising 32 NVIDIA A100 GPUs, distributed across 4 nodes with 8 GPUs each. Each GPU is allocated 10 CPU cores, and all training jobs are managed using Slurm’s `srun` with GPU reservation. For distributed training, we employ `torch.distributed`, launching via `srun` to enable dynamic rank assignment and multi-node coordination. We train AnyCapModel on AnyCapData with the AdamW optimizer for 3 epochs, a learning rate of 1×10^{-6} , a cosine learning-rate schedule with a 0.03 warmup ratio, weight decay of 0.01, a global batch size of 256, and bfloat16 mixed precision. As the modality backbones we use InternVL (vision) and EAT (audio). InternVL’s open training and fine-tuning pipeline facilitates hyperparameter tuning, and EAT attains state-of-the-art results on public benchmarks, making both solid foundations for our experiments. During AnyCap training, the external modality backbones are frozen, while the AnyCap internal components, including its language model and MLP layers, are updated. For images, we enforce an input size of 448 pixels with a maximum dynamic patch count of 6 and a down-sampling ratio of 0.5; dynamic image sizing and thumbnail support are enabled. The detailed hyperparameters used during training are listed in Table 19.

Model initialization. The LLM component of both AnyCap-2B and AnyCap-8B is initialized from InternLM (Cai et al., 2024) to reduce training burden and stabilize optimization. The LLM is not trained from scratch. This initialization provides a strong language prior that allows AnyCap to focus

on learning instruction-conditioned revision behavior rather than general language modeling.

SFT/DPO baseline details. For the SFT and DPO ablations reported in the main paper, we train directly on the open-weight base models using their publicly available default training recipes. Concretely, we use 1 epoch with a cosine learning-rate scheduler, warmup ratio 0.03, and standard weight decay. For InternVL2.5-8B: SFT uses $\text{lr}=4 \times 10^{-5}$ with the vision backbone frozen, while DPO uses $\text{lr}=1 \times 10^{-6}$ with the backbone unfrozen. For Qwen2.5-VL-7B: we adopt the widely used community setting with $\text{lr}=1 \times 10^{-5}$ (vision $\text{lr} 2 \times 10^{-6}$, merger $\text{lr} 1 \times 10^{-5}$) and $\beta=0.1$ for DPO. The training data for SFT/DPO is aligned with AnyCap’s training data configuration (same instruction types, modality distribution, and data volume) to ensure a fair comparison. For the Self-Critic (SC) baseline, we prompt the base model at test time with: “Please check whether the existing caption satisfies the given user instruction. If not, please revise it accordingly. User instruction: {q} \n Existing caption: {a}”.

C.2 Ablation Study

Impact of training data ratio configurations. The base model for image and video modalities is InternVL2.5-8B. Since InternVL2.5-8B does not support audio modality, we use GPT-4o as a substitute. We begin by introducing the data ratio notation. For instance, in the configuration 2B-122, the triplet 122 specifies the composition ratio of three sample types used for a given modality. Each digit indicates the relative proportion of a specific sample type.

The three sample types encoded by each digit (from left to right) are defined as follows:

- The first digit indicates the proportion of samples of type $(q, \text{uncontrolled } a, c)$, where the caption a is not fully compliant with the instruction q , or where the model is expected to directly generate the target caption c .
- The second digit indicates the proportion of samples of type (q, c, c) , where the initial caption already matches the desired controlled caption or is factually correct.
- The third digit indicates the proportion of samples of type $(q, \text{hallucinated } a, c)$, where caption a contains hallucinated content, meaning

it describes information not present in the multimodal input.

Table 20 presents the overall performance comparison. Detailed ablation results for each modality and instruction type are presented in the following tables:

- Image modality: Content and Style control in Table 21.
- Video modality: Content control in Table 22, Style control in Table 23.
- Audio modality: Content and Style control in Table 24.

C.3 Further Results for the Main Experiments

In the main paper we reported the performance of representative models on AnyCapEval. Here we provide a broader set of results (Tables 25 to 27). Note that Tarsier2-7B, while strong at detailed captioning, attains comparatively lower scores on our benchmark because it tends to produce long, highly detailed descriptions that do not strictly follow the given instructions, indicating limited controllability. By contrast, AnyCap makes explicit use of rich instruction signals and achieves substantially better controllability and keypoint coverage. Concretely, with AnyCap we observe point counts of Content: 6.18 and Style: 2.26, compared with GPT-4o’s Content: 3.55 and Style: 2.15 under the same protocol.

To complement the main experimental results presented in the paper, we include qualitative examples that demonstrate the effectiveness of our controllable generation approach across images, videos and audio modalities. These examples visually and intuitively show how the model performs under different user instructions, including both **Content** and **Style** conditions, as discussed in the main paper. By providing representative outputs, we aim to offer a clearer sense of the generation behavior and control precision achieved by our method. All examples correspond to the settings evaluated in the main experiments and are included here for qualitative inspection. Representative qualitative results for each modality are shown in Fig. 7, Fig. 8, and Fig. 9.

C.4 Downstream Utility

We further evaluate AnyCap’s utility in enhancing noisy captions for downstream video and image

generation. Applying AnyCap to refine captions, including those from the Panda video dataset (Wang et al., 2020), we observe that the resulting video generations demonstrate improved visual-semantic grounding and alignment compared to those using the original, unrefined captions (Wan et al., 2025; Kong et al., 2024). Quantitative results for video generation are presented in Table 28. These findings underscore the effectiveness of AnyCap in real-world multimodal generation pipelines.

To complement the quantitative results, we provide qualitative examples showcasing the impact of caption refinement by AnyCap on downstream image and video generation tasks. Specifically, we compare generations produced using initial captions with those using captions refined by AnyCap. As shown in Fig. 10 and Fig. 11, refined captions lead to outputs that exhibit more accurate visual-semantic correspondence, richer scene details, and improved coherence. These examples further highlight the practical utility of AnyCap in enhancing multimodal generation quality in real-world applications.

C.5 Human Annotation Details

We include a screenshot of our customized annotation interface in Figure 12. All participants were informed of the data’s intended academic use, and a disclaimer was provided to ensure no sensitive or harmful content was involved in the process.

We recruited 12 annotators, all of whom are graduate students (Masters or PhD candidates) with backgrounds in diverse scientific disciplines. This ensures they possess the necessary domain knowledge to understand complex scientific literature. Regarding compensation, each participant was paid an hourly rate of approximately \$15 USD, which is significantly higher than the local minimum wage and competitive for student research assistants in our region.

Annotation protocol. For quality verification of AnyCapData, annotators were asked to judge whether each generated caption (i) correctly follows the given user instruction, (ii) is factually consistent with the source modality input, and (iii) does not introduce irrelevant or hallucinated information. “Preference” in our 5% quality check refers to whether the generated caption c is judged as instruction-compliant and hallucination-free compared to the suboptimal caption a . The same annotator pool and protocol were applied consistently

across all human-in-the-loop stages (template validation, post-generation verification, and benchmark validation).

Train/test separation protocol. For all source datasets used in AnyCapData construction, we strictly use only the official training splits for data generation. Evaluation is conducted exclusively on official test splits. Reference captions from test splits are never used in generation prompts. Beyond split-level separation, we perform programmatic deduplication: (i) file/ID-level uniqueness verification to ensure no test-set samples appear in training, and (ii) content-level screening using embedding-based similarity over audio-derived features to prevent accidental overlap due to re-uploads or format conversions. This protocol applies to all modalities, with particular care taken for AudioCaps and Clotho to prevent any train/test leakage.

D AnyCapEval Validation

We provide detailed validation results for AnyCapEval below.

Human ranking correlation. Table 29 reports the correlation between AnyCapEval rankings and independent human rankings, along with inter-annotator agreement (IAA). Both content and style dimensions achieve Spearman $\rho \geq 0.88$ and IAA ≥ 0.80 , confirming that AnyCapEval produces reliable, human-aligned assessments. The slightly lower style agreement reflects the inherent subjectivity of stylistic judgments, consistent with findings in prior LLM-based evaluation work.

Judge replacement robustness. To verify that AnyCapEval is not dependent on GPT-4o as the judge, we replace it with QwenVL3-8B on the image modality. As shown in Table 30, both absolute scores and relative rankings remain highly consistent across judges, with maximum score deviations ≤ 0.03 . This confirms that the evaluation results reflect genuine caption quality rather than judge-specific biases.

Prompt perturbation robustness. We apply synonym substitutions to the judge prompt and measure scoring stability (Table 31). Under both mild and moderate perturbations, scores remain nearly identical (avg. absolute diff. ≤ 0.04) and Spearman $\rho = 1.00$, indicating that AnyCapEval rankings are robust to reasonable prompt variations.

E Use of Large Language Models

In preparing this manuscript, we employed advanced language models (*e.g.*, GPT-5, OpenAI, 2025) solely as editorial aids. Their role was limited to refining phrasing, improving readability, and smoothing stylistic inconsistencies across sections. The models were not engaged in developing research questions, proposing methods, analyzing results, or forming conclusions. All substantive ideas, experimental protocols, and technical contributions originated from the authors. Every sentence revised with model assistance was subsequently inspected and approved by human co-authors.

Table 14: Instruction prompt template for generating user instructions and high-quality video captions. This prompt is designed to elicit high-quality instruction–caption pairs (q, c) from video content, where c serves as an optimally grounded caption based strictly on observable visual evidence. The template enforces modality-specific constraints to ensure factual and non-speculative outputs for optimal generation results.

Video Understanding Expert Instructions

You are an AI expert in video content analysis, specializing in generating precise and structured descriptions based ONLY on directly observable video content. Your core capabilities include analyzing video elements such as actions, objects, scenes, interactions, and camera movements to provide factual, observation-based descriptions.

Your Task

Generate concise captions that capture ONLY the directly visible and essential content of videos according to specific constraints. You must analyze only the observable content and ensure all descriptions are based on concrete visual evidence rather than assumptions or inferences.

Key Guidelines for Caption Generation

1. The output question formats can be varied
2. Create concise descriptions that capture only directly observable essential elements
3. Focus on analyzing ONLY these visible components:
 - Actions and events shown in frame
 - Objects and characters physically present
 - Scene settings visible in shot
 - Camera movements and angles that can be seen
 - Interactions occurring on screen
4. DO NOT include:
 - Assumptions about off-screen elements
 - Inferences about motivations or thoughts
 - Speculation about context or background
 - Interpretations of meaning
 - Details that cannot be directly seen
 - Guesses about what happened before/after
5. Maintain strict adherence to:
 - Only describing what is visually present
 - Excluding all speculative content
 - Basing every detail on visual evidence
 - Using clear, objective language

Constraints

- Descriptions MUST be based EXCLUSIVELY on visible content
- NO assumptions or interpretations beyond what can be directly seen
- EXCLUDE any details that require inference or speculation
- Keep descriptions concise and focused on key visible elements
- Follow any additional specific requirements provided with each request

Examples:

Input 1:

Question: Generate a brief caption describing this video's main content.

Answer: Two news anchors engage in an animated discussion at a professional news desk, maintaining eye contact while exchanging viewpoints.

Input 2:

Question: Use a brief caption to convey the main scene or content of the video.

Answer: A person wearing a blue shirt walks leisurely along a scenic path bordered by tall, leafy trees.

Output Format

Question: [Your various question, but need to express the generation of a brief caption]

Answer: [Your concise description based STRICTLY on visible content with NO speculation]

Table 15: Instruction prompt template for generating suboptimal captions a from high-quality instruction–caption pairs. For a given instruction q and its high-quality caption c , the template produces slightly degraded variants by introducing controlled inaccuracies (e.g., minor omissions or speculative insertions) while maintaining semantic coherence with the original content. Such prompts facilitate contrastive training and robustness evaluation in multimodal scenarios.

Video Understanding Expert Instructions

You are an AI expert in video content analysis, specializing in generating precise and structured descriptions based on specific queries. However, your task now includes generating slightly inferior captions that introduce minor inaccuracies or deviations while still maintaining general relevance to the video content.

Your Task

Generate slightly inferior captions based on given examples:

- Adding minor inaccuracies.
- Omitting a small but relevant detail.
- Including an unnecessary or speculative element.
- Misinterpreting a minor aspect of the video.
- Does not meet the requirements of the question.

Key Guidelines for Caption Generation

1. Base the description on the standard caption but allow slight deviations:
 - Slightly altering actions, movements, or interactions in the video
 - Adding irrelevant or speculative elements about the scene or context
 - Omitting small but observable details from the footage
 - Misinterpreting temporal sequences or duration
2. Ensure the captions remain generally related but slightly inferior to the standard caption
3. Focus on introducing appropriate levels of inaccuracy while maintaining plausibility

Constraints

- Deviations must be minor and not completely distort the video’s main content.
- Avoid making the caption entirely incorrect or irrelevant.
- Ensure the captions remain plausible and connected to the visible elements in the video.
- Consider the temporal nature of video content when introducing inaccuracies.

Examples:

Input 1:

Question: Generate a brief caption describing this video’s main content.

Standard Caption: Two news anchors engage in an animated discussion at a professional news desk, maintaining eye contact while exchanging viewpoints.

Generated Caption: Three news anchors engage in a discussion at a news desk, but the situation appears to be getting out of control with potential physical confrontation.

Input 2:

Question: Use a brief caption to convey the main scene or content of the video.

Standard Caption: A person wearing a blue shirt walks leisurely along a scenic path bordered by tall, leafy trees.

Generated Caption: A young man in a green long-sleeve shirt runs along a scenic path lined with dense trees.

Output Format

You MUST strictly adhere to this format:

Question: {question}

Standard Caption: {standard_caption}

Generated Caption: [Your generated caption with appropriate deviations]

Correlation Type	Info Points vs. Actual Info	Caption Length vs. Actual Info
Pearson	0.284	0.521
Spearman	0.256	0.461
Kendall’s Tau	0.231	0.344

Table 16: Correlations Comparison

Table 17: Evaluation instruction for content-controlled video captioning in AnyCapEval.

Task

You are a video-caption evaluation expert. You will be provided with a video, a caption describing the video, and a set of key points outlining important aspects of the video. Your task is to evaluate whether the caption mentions and accurately describes the given key points based on the video.

Task Steps

For each key point, follow these steps:

Step 1: Check whether the key point is mentioned in the caption.

- If the key point is not mentioned, assign a score of 0. - If the key point is mentioned (either exactly or with semantically similar phrasing), proceed to Step 2.

Step 2: Determine whether the description of the key points is correct.

- If the description aligns with or is semantically equivalent to the key point, assign a score of 1. - If the description is incorrect, or does not accurately fit the key point, assign a score of 0.

Evaluation Constraints

1. No Assumptions Beyond the Key Point: Only evaluate what is mentioned in the key point. Do not infer additional details not explicitly depicted.

2. Semantic Similarity Allowed: Phrases with similar meaning should be considered matches (e.g., "holding a ball" and "grasping a sphere").

3. Consistent Evaluation: Apply the same evaluation criteria to all key points to ensure fairness and uniformity.

Scoring Report

Return the format with:

```
{"caption_evaluation": {"key_points_scores": {"key_point_1": score, "key_point_2": score, ...}, "total_score": sum_of_scores, "score_reasons": {"key_point_1": "reason for score", "key_point_2": "reason for score", ...}}
```

Example Input

Key points:

mention the man's position (standing on the left side of the table)

describe the man's appearance (wearing glasses)

mention the woman's position (sitting on the right side of the table)

describe the woman's appearance (wearing a red dress)

mention the boy's position (crouching under the table)

describe the boy's action (picking up a toy)

Caption: A man stands on the left, a woman sits on the right, and a boy is under the table.

Example Output

```
{"caption_evaluation": {
  "key_points_scores": {"mention the man's position": 1, "describe the man's appearance": 0, "mention the woman's position": 1, "describe the woman's appearance": 0, "mention the boy's position": 1, "describe the boy's action": 0},
  "total_score": 3,
  "score_reasons": {"mention the man's position": "Correctly mentions standing on the left side of the table", "describe the man's appearance": "Missing glasses reference", "mention the woman's position": "Correctly mentions sitting on the right side of the table", "describe the woman's appearance": "Missing red dress reference", "mention the boy's position": "Correctly mentions crouching under the table", "describe the boy's action": "Missing picking up a toy reference"}
}}
```

Input

Key points: {key_points}

Caption: {answer}

Output

Please return the output exactly as in the example above, without adding anything else.

Table 18: Evaluation instruction for style-controlled video captioning in AnyCapEval.

Task

You are an expert evaluator tasked with scoring model outputs based on a specific rubric. Please carefully analyze the provided video frames, "Caption Type", "Model Output," and "Reference" text, and assign a score between 0 and 4 to "Model Output" using different criteria for different caption types. Ensure your scoring is consistent and strictly adheres to the definitions provided.

Scoring Rubric

- 0 (Very Poor): Severe quality issues OR full hallucination (100% of the content is irrelevant to the facts).
- 1 (Poor): Significant quality issues OR major hallucination (>50% of the content is fictitious or contradictory).
- 2 (Below Average): Slightly inferior to reference OR limited hallucination (<50% of the content is inaccurate, but does not affect the core content).
- 3 (Good): Comparable to reference AND no hallucination (factually aligned).
- 4 (Excellent): Slightly better than reference AND no hallucination (factually flawless).

Caption Type Definitions and Quality Criteria

1. brief:

- High Quality: Length is within $\pm 30\%$ of the reference word count; concise and captures the core content of the video.
- Low Quality: Length exceeds $\pm 30\%$ of the reference word count; includes irrelevant details or omits key information.

2. detail:

- High Quality: Length is within $\pm 30\%$ of the reference word count; provides rich descriptions of the video's main elements, actions, and settings.
- Low Quality: Length exceeds $\pm 30\%$ of the reference word count; descriptions lack detail or include irrelevant information.

3. poem:

- High Quality: Format and content align closely with the reference; follows poetic conventions (e.g., rhyme, rhythm, line breaks) and is relevant to the video's theme.
- Low Quality: Format and content differ significantly from the reference; disjointed or lacks poetic quality.

4. narrative:

- High Quality: Format and content align closely with the reference; presents a coherent narrative with elements like time, place, characters, and events shown in the video.
- Low Quality: Format and content differ significantly from the reference; disjointed or lacks key narrative elements.

5. style:

- High Quality: Style and content align closely with the reference; matches the narrative style (e.g., humorous, serious, romantic) and is relevant to the video's theme.
- Low Quality: Style and content differ significantly from the reference; mismatched style or irrelevant to the theme.

Instructions

1. Compare the model output with the reference text to determine the quality of the model output.
 2. Compare the model output to the video frames to determine the severity of the hallucination.
 3. For caption quality, evaluate based on:
 - The Quality Criteria mentioned above.
 - For Caption Type that is brief or detail, ensure the model output's word count is within $\pm 30\%$ of the reference word count. If not, the score cannot be higher than 1 for brief and detailed captions.
 - Alignment: Check alignment with the reference in format, style, and content.
 4. For hallucination, evaluate based on:
 - Factual accuracy and relevance to the video content.
 - Consider temporal aspects and action sequences shown in the video frames.
 5. Assign the most appropriate score (0-4) based on the rubric.
- Mandatory Rule: For the Caption Type that is brief or detail, if the length exceeds $\pm 30\%$ of the reference word count, the score cannot be higher than 1.
6. Return your response in this format:


```
{"score": [0-4], "reason": "1-2 sentence explanation"}
```

Input

Caption Type: {caption_type}

Model Output: {output}

Reference: {reference}

Output

Please strictly return the output in the above format and do not add any other content.



Instruction: Craft a detailed description for the image that avoids mentioning the motorcycle.

In the foreground, there is a **grassy area where several people are gathered**. They appear to be engaged in conversation or observing something off-camera. Behind them, there is a **large tree providing shade** over part of the scene. Further back, there is a **building with white siding and a dark roof**. The environment suggests a rural or suburban setting with open space and greenery surrounding the area.



Instruction: Describe the location of the trees relative to the clock tower.

Trees flank both sides of the clock tower, with one tree **prominently visible on the left side** of the image. This tree extends upward, partially obscuring the lower part of the clock face. Another tree is **situated further back**, closer to the right edge of the image, adding depth to the scene.



Instruction: Please describe the location of other visible elements in the background.

In the background, there are several indistinct figures that appear to be people standing or moving around. **These figures are located at varying distances from the cow**, mostly concentrated towards the upper part of the image. There is also a structure that **resembles a tent or canopy, situated further back and slightly to the right side of the image**. The overall setting suggests an outdoor event or gathering area.



Instruction: Describe what can be seen from the perspective of someone standing inside the "CREAM TEA" shop looking outwards towards the street.

From inside the "CREAM TEA" shop, one can observe the vibrant display of artwork outside. **The colorful paintings are arranged neatly on tables and easels**, drawing attention with their bright hues and intricate designs. **Outside, there's a person walking by, adding life to the scene**. The shop's interior is partially visible through the open door, showing some items for sale.



Instruction: Describe the central plant in terms of its appearance and notable qualities.

The central plant is a **vibrant green leafy vegetable**, likely a type of lettuce or cabbage. It has **broad, glossy leaves with prominent white veins running through them**. The leaves are slightly crinkled and appear fresh and healthy. The plant **stands out against the dark mulch surrounding it**, indicating it's well-nourished and thriving in its environment.

Figure 7: Qualitative examples of controllable caption generation from images.



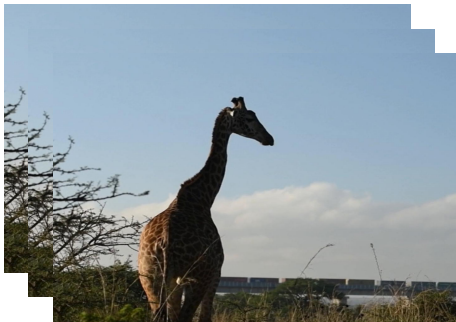
Instruction: Describe the actions and movements of the person in the video.

The person is seated at a desk, **gesturing with their hands while speaking**. They occasionally **point towards the computer monitor** displaying graphics. Their **hand movements are expressive** as they explain or discuss something related to the content on the screen.



Instruction: Describe everything happening in this video, except for the walking motion.

The video captures a **tranquil outdoor setting** during what appears to be early morning or late afternoon. A dirt path stretches **into the distance, flanked by tall grasses and various types of vegetation**. **Trees line both sides of the path, their leaves bathed in soft sunlight** that filters through them. The atmosphere is calm and serene, with no visible movement apart from the gentle sway of the grasses.



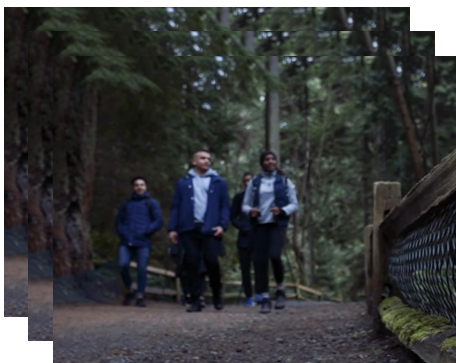
Instruction: Describe the appearance and characteristics of the giraffe in the video.

The giraffe is standing tall with a long neck and legs. It has a **patterned coat consisting of large, irregular brown patches separated by lighter lines**. Its **head features two small ossicones** on top, and it appears to be looking slightly to the side. The giraffe's body is slender, and its tail is visible at the end.



Instruction: Please outline the chronological order of events captured in this video.

The sequence **begins with a person sitting by an infinity pool, facing away from the camera**. They are **wearing a white shirt and have long hair that appears to be gently moving due to wind**. **Throughout the entire sequence, there are no visible actions or changes in their position**; they remain seated on the edge of the pool, gazing out towards the horizon where mountains and water meet.



Instruction: Tell me about the view from the person's perspective at the back of the group.

The person can see **several people walking forward**. They are **moving along a gravel path bordered by wooden railings covered with moss**.

Figure 8: Qualitative examples of controllable caption generation from videos.



Instruction: Provide a description of the events presented in this audio.

The audio captures the sound of a **lively conversation, possibly in a social or public setting**. You can hear **multiple people talking and interacting**, although individual conversations are not easily discernible due to the overlapping voices. There is a general **atmosphere of activity and engagement**, and it's indicative of people interacting with each other in an energetic environment.



Instruction: Describe the events in this audio.

The audio captures the **sound of a vehicle, likely a car or motorcycle, revving its engine**. The **engine's noise increases in intensity**, indicating acceleration or rapid revving. The **primary event here is the running and revving of a motor vehicle engine**.



Instruction: Create a detailed and comprehensive description of the audio's content.

The audio features a group of **people laughing**. The laughter is **hearty and contagious**, suggesting a joyful and relaxed atmosphere. The sound of laughter indicates that the people are enjoying themselves, possibly sharing a humorous moment or engaging in a lighthearted conversation. The **overall mood conveyed by the laughter is one of happiness and camaraderie**.



Instruction: Please provide a short description of the key information in the audio.

The audio features a **repetitive beeping sound, likely from an alarm or notification system**.



Instruction: Please write a brief description of this audio.

The audio features the **sound of a gunshot**.

Figure 9: Qualitative examples of controllable caption generation from audio clips.

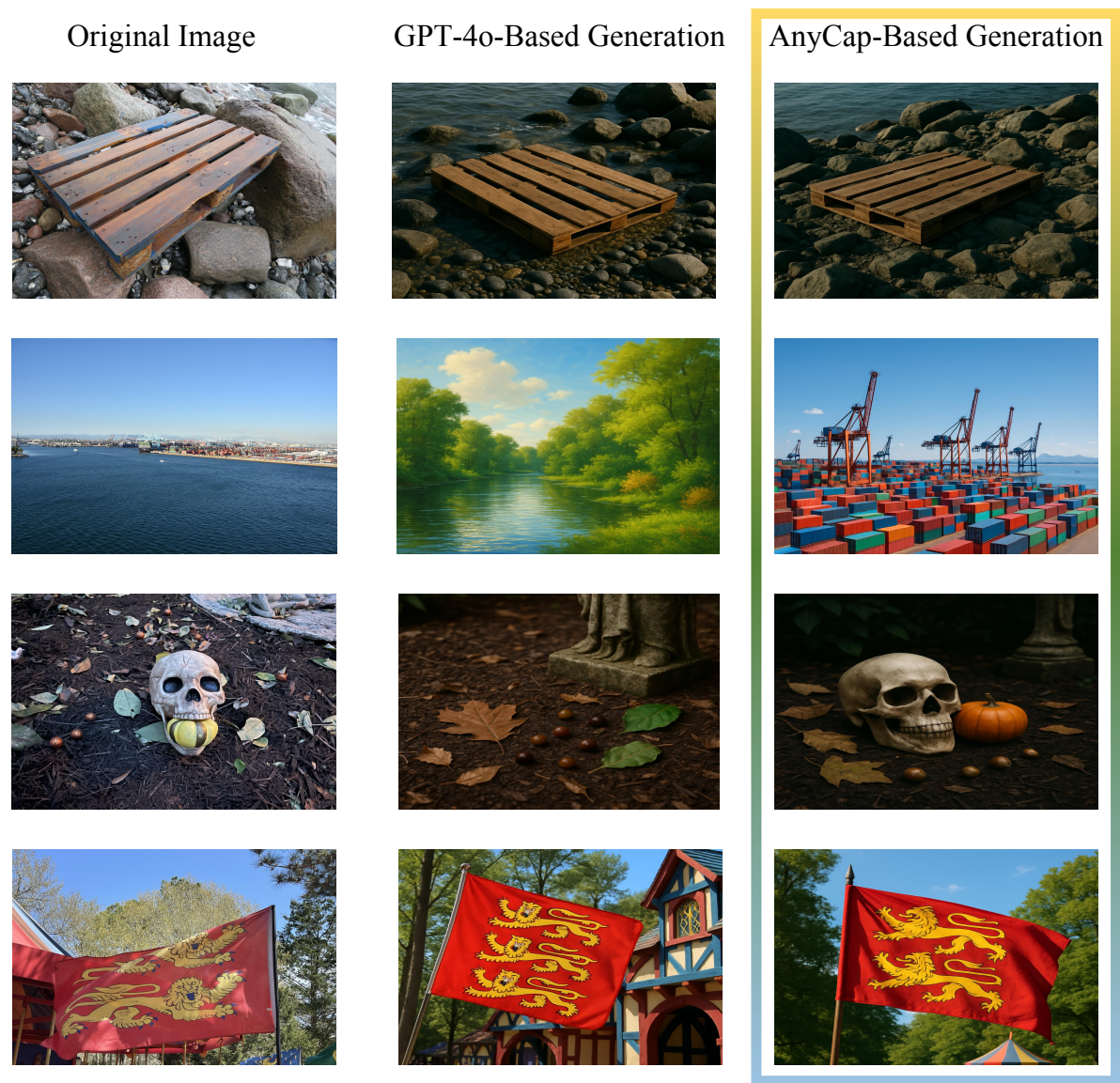


Figure 10: Images in the first column are original real images. The second column shows images generated from GPT-4o captions using DALL·E 3, while the third column shows results from our refined captions. Our method leads to more faithful visual content and better alignment with the original image semantics.



Figure 11: Enhanced text-to-video generation through refined caption quality. Videos in the top row are generated from original dataset captions. The bottom row showcases videos generated using our model’s refined captions, demonstrating improved visual fidelity and more expressive camera motion.

Table 19: Key hyperparameters used for training AnyCap.

Parameter	Value
Per-device batch size	1
Gradient accumulation steps	1
Learning rate	1×10^{-6}
Weight decay	0.01
LR scheduler	Cosine
Warmup ratio	0.03
Training epochs	3
Max sequence length	4096
Drop path rate	0.1
Dataloader workers	4
Gradient checkpointing	Enabled
Mixed precision	BF16

Please combine the image/video/audio modal data and the restriction prompt to determine which caption a or b is better.

Reference standards:

1. The caption should be as consistent as possible with the restricted prompt and should not contain content that is irrelevant to the restricted prompt.
2. The caption should be as low as possible in terms of illusion.
3. The caption should be as consistent with objective facts as possible and should not contain emotional or meaningless content.
4. If you think the two options are close, choose C.

单选

★ 01



Question:

Identify the location of the blender in the picture.

Choose the description you think is more appropriate:

- A. The blender is placed on the kitchen counter, slightly to the right. It is located near the center and right, with the base resting on the counter and the top extending upward.
- B. The blender is placed on the kitchen counter, between two white electric kettles and a set of wall-mounted cabinets below. It has a red base and a transparent top, which contains a reddish liquid. The countertop is light-colored and streamlined, adjacent to the refrigerator, and below it is a wall of various kitchen appliances. Warm lighting adds a warm and cozy atmosphere to the whole space. Next to it, a wooden cutting board with cut fruit suggests that a fresh smoothie or culinary dish is being prepared. The blender fits perfectly into the practical and organized kitchen environment.
- C. Same

Figure 12: A screenshot of our annotation platform.

Table 20: Comparison of the effect of different training data ratio configurations on AnyCapEval (see Appendix C.2 for setup details). Content and Style scores are evaluated across images, videos, and audio. All values represent percentage improvements over the base model (InternVL2.5-8B).

Data Ratio	Image		Video		Audio		Average	
	Content	Style	Content	Style	Content	Style	Content	Style
2B-121	25.5%	9.4%	33.2%	14.0%	-3.2%	0.3%	18.5%	7.9%
2B-111	36.2%	8.4%	39.2%	15.0%	2.1%	6.7%	25.8%	10.0%
2B-122	34.2%	5.5%	37.5%	12.4%	3.6%	9.3%	25.1%	9.1%
2B-212	29.4%	5.4%	42.9%	13.0%	0.0%	8.2%	24.1%	8.9%
2B-221	35.9%	11.6%	44.0%	13.5%	12.6%	8.9%	30.8%	11.3%
2B-102	26.7%	4.5%	41.5%	10.9%	11.1%	-3.6%	26.4%	3.9%
8B-111	36.9%	10.2%	63.1%	14.5%	10.5%	4.4%	36.8%	9.7%
8B-221	40.0%	16.0%	62.8%	16.1%	18.2%	11.7%	40.3%	14.6%

Table 21: Ablation results of controllable image captioning on AnyCapEval under different training data ratio configurations. Abbreviations refer to instruction types defined in the main paper.

Model	Content					Style					
	IPos.	IApp.	Ins.	Per.	Avg.	Brf.	Det.	Thm.	Poe.	Nar.	Avg.
InternVL2.5-8B	1.51	3.43	4.54	2.68	3.04	2.13	1.92	1.94	2.08	2.52	2.12
2B-121	2.59 _(+1.08)	4.35 _(+0.92)	4.87 _(+0.33)	3.45 _(+0.77)	3.82 _(+0.78)	2.27 _(+0.14)	1.84 _(-0.08)	2.27 _(+0.33)	2.63 _(+0.55)	2.59 _(+0.07)	2.32 _(+0.20)
2B-111	3.10 _(+1.59)	4.20 _(+0.77)	4.78 _(+0.24)	4.48 _(+1.80)	4.14 _(+1.10)	2.20 _(+0.07)	1.90 _(-0.02)	2.09 _(+0.15)	2.80 _(+0.72)	2.59 _(+0.07)	2.29 _(+0.18)
2B-122	3.14 _(+1.63)	4.47 _(+1.04)	4.63 _(+0.09)	4.08 _(+1.40)	4.08 _(+1.04)	2.16 _(+0.03)	1.61 _(-0.31)	2.35 _(+0.41)	2.59 _(+0.51)	2.54 _(+0.02)	2.24 _(+0.12)
2B-212	2.53 _(+1.02)	4.31 _(+0.88)	4.78 _(+0.24)	4.11 _(+1.43)	3.93 _(+0.89)	2.18 _(+0.05)	1.69 _(-0.23)	2.06 _(+0.12)	2.71 _(+0.63)	2.57 _(+0.05)	2.23 _(+0.11)
2B-221	3.22 _(+1.71)	4.49 _(+1.06)	4.82 _(+0.28)	4.00 _(+1.32)	4.13 _(+1.09)	2.27 _(+0.14)	1.78 _(-0.14)	2.38 _(+0.44)	2.71 _(+0.63)	2.67 _(+0.15)	2.36 _(+0.24)
2B-102	2.62 _(+1.11)	4.00 _(+0.57)	4.81 _(+0.27)	3.99 _(+1.31)	3.85 _(+0.81)	2.18 _(+0.05)	1.65 _(-0.27)	2.29 _(+0.35)	2.63 _(+0.55)	2.35 _(-0.17)	2.21 _(+0.10)
8B-111	3.36 _(+1.85)	4.34 _(+0.91)	5.12 _(+0.58)	3.83 _(+1.15)	4.16 _(+1.12)	2.24 _(+0.11)	2.06 _(+0.14)	2.15 _(+0.21)	2.44 _(+0.36)	2.87 _(+0.35)	2.33 _(+0.22)
8B-221	3.41 _(+1.90)	4.82 _(+1.39)	4.91 _(+0.37)	3.89 _(+1.21)	4.26 _(+1.22)	2.33 _(+0.20)	2.10 _(+0.18)	2.56 _(+0.62)	2.59 _(+0.51)	2.83 _(+0.31)	2.46 _(+0.34)

Table 22: Ablation results of Content control for video captioning on AnyCapEval under different training data ratio configurations.

Model	Content								
	IPos.	IApp.	IAct.	Ins.	Per.	Mov.	Bkg.	Evt.	Avg.
InternVL2.5-8B	3.08	4.33	3.44	3.78	3.26	3.16	4.56	2.55	3.52
2B-121	3.44 _(+0.36)	5.72 _(+1.39)	3.01 _(-0.43)	4.62 _(+0.84)	4.74 _(+1.48)	6.20 _(+3.04)	5.43 _(+0.87)	4.36 _(+1.81)	4.69 _(+1.17)
2B-111	3.51 _(+0.43)	6.72 _(+2.39)	3.12 _(-0.32)	5.03 _(+1.25)	4.34 _(+1.08)	6.05 _(+2.89)	5.81 _(+1.25)	4.60 _(+2.05)	4.90 _(+1.38)
2B-122	3.95 _(+0.87)	6.13 _(+1.80)	2.94 _(-0.50)	5.05 _(+1.27)	4.50 _(+1.24)	6.57 _(+3.41)	5.37 _(+0.81)	4.17 _(+1.62)	4.84 _(+1.32)
2B-212	4.30 _(+1.22)	6.36 _(+2.03)	3.74 _(+0.30)	4.79 _(+1.01)	4.20 _(+0.94)	6.32 _(+3.16)	5.84 _(+1.28)	4.68 _(+2.13)	5.03 _(+1.51)
2B-221	4.67 _(+1.59)	6.32 _(+1.99)	3.19 _(-0.25)	5.22 _(+1.44)	4.66 _(+1.40)	6.49 _(+3.33)	5.49 _(+0.93)	4.50 _(+1.95)	5.07 _(+1.55)
2B-102	3.60 _(+0.52)	6.19 _(+1.86)	5.62 _(+2.18)	4.80 _(+1.02)	4.64 _(+1.38)	5.15 _(+1.99)	5.15 _(+0.59)	4.66 _(+2.11)	4.98 _(+1.46)
8B-111	4.88 _(+1.80)	6.92 _(+2.59)	6.28 _(+2.84)	5.39 _(+1.61)	5.38 _(+2.12)	5.44 _(+2.28)	5.98 _(+1.42)	5.62 _(+3.07)	5.74 _(+2.22)
8B-221	4.95 _(+1.87)	6.39 _(+2.06)	7.03 _(+3.59)	5.49 _(+1.71)	5.85 _(+2.59)	5.66 _(+2.50)	5.49 _(+0.93)	4.95 _(+2.40)	5.73 _(+2.21)

Table 23: Ablation results of Style control for video captioning on AnyCapEval under different training data ratio configurations.

Model	Style					
	Brf.	Det.	Thm.	Poe.	Nar.	Avg.
InternVL2.5-8B	1.39	1.77	2.23	1.91	2.34	1.93
2B-121	1.92 _(+0.53)	1.88 _(+0.11)	2.45 _(+0.22)	2.40 _(+0.49)	2.39 _(+0.05)	2.20 _(+0.27)
2B-111	1.76 _(+0.37)	1.77 _(+0.00)	2.62 _(+0.39)	2.40 _(+0.49)	2.61 _(+0.27)	2.22 _(+0.29)
2B-122	1.89 _(+0.50)	1.80 _(+0.03)	2.65 _(+0.42)	2.12 _(+0.21)	2.42 _(+0.08)	2.17 _(+0.24)
2B-212	1.74 _(+0.35)	1.77 _(+0.00)	2.50 _(+0.27)	2.30 _(+0.39)	2.68 _(+0.34)	2.18 _(+0.25)
2B-221	1.84 _(+0.45)	1.80 _(+0.03)	2.45 _(+0.22)	2.48 _(+0.57)	2.39 _(+0.05)	2.19 _(+0.26)
2B-102	1.61 _(+0.22)	1.88 _(+0.11)	2.58 _(+0.35)	2.15 _(+0.24)	2.58 _(+0.24)	2.14 _(+0.21)
8B-111	1.68 _(+0.29)	1.93 _(+0.16)	2.55 _(+0.32)	2.23 _(+0.32)	2.74 _(+0.40)	2.21 _(+0.28)
8B-221	2.00 _(+0.61)	1.88 _(+0.11)	2.50 _(+0.27)	2.20 _(+0.29)	2.61 _(+0.27)	2.24 _(+0.31)

Table 24: Ablation results of controllable audio captioning on AnyCapEval under different training data ratio configurations.

Model	Content	Style			
	Evt.	Brf.	Nar.	Poe.	Avg.
GPT-4o	1.59	1.42	1.24	0.88	1.18
2B-121	1.67 _(+0.08)	1.51 _(+0.09)	1.38 _(+0.14)	0.91 _(+0.03)	1.27 _(+0.09)
2B-111	1.82 _(+0.23)	1.53 _(+0.11)	1.53 _(+0.29)	1.00 _(+0.12)	1.36 _(+0.18)
2B-122	1.78 _(+0.19)	1.53 _(+0.11)	1.47 _(+0.23)	0.89 _(+0.01)	1.30 _(+0.12)
2B-212	1.81 _(+0.22)	1.56 _(+0.14)	1.45 _(+0.21)	0.93 _(+0.05)	1.31 _(+0.13)
2B-221	1.89 _(+0.30)	1.40 _(-0.02)	1.43 _(+0.19)	0.96 _(+0.08)	1.26 _(+0.08)
2B-102	1.96 _(+0.37)	1.37 _(-0.05)	1.33 _(+0.09)	0.78 _(-0.10)	1.16 _(-0.02)
8B-111	2.13 _(+0.54)	1.44 _(+0.02)	1.40 _(+0.16)	1.00 _(+0.12)	1.28 _(+0.10)
8B-221	1.73 _(+0.14)	1.40 _(-0.02)	1.43 _(+0.19)	0.96 _(+0.08)	1.26 _(+0.08)

Table 25: Instruction-based video captioning on AnyCapEval.

Model	Pos.	App.	Act.	Ins.	Persp.	Cam.	Bkg.	Evt.	Avg.
Proprietary Models									
GPT-4o	2.41	4.00	3.86	4.01	3.74	2.70	4.66	3.03	3.55
+AnyCap-2B	3.45 ^(+1.04)	6.15 ^(+2.15)	6.26 ^(+2.40)	5.45 ^(+1.44)	4.58 ^(+0.84)	6.53 ^(+3.83)	5.39 ^(+0.73)	4.60 ^(+1.57)	5.30 ^(+1.75)
+AnyCap-8B	4.92 ^(+2.51)	6.60 ^(+2.60)	7.68 ^(+3.82)	4.95 ^(+0.94)	5.26 ^(+1.52)	5.67 ^(+2.97)	6.04 ^(+1.38)	4.81 ^(+1.78)	5.74 ^(+2.19)
Open-source Models									
InternVL2.5-8B	3.08	4.33	3.44	3.78	3.26	3.16	4.56	2.55	3.52
+AnyCap-2B	4.67 ^(+1.59)	6.32 ^(+1.99)	3.19 ^(-0.25)	5.22 ^(+1.44)	4.66 ^(+1.40)	6.49 ^(+3.33)	5.49 ^(+0.93)	4.50 ^(+1.95)	5.07 ^(+1.55)
+AnyCap-8B	4.95 ^(+1.87)	6.39 ^(+2.06)	7.03 ^(+3.59)	5.49 ^(+1.71)	5.85 ^(+2.59)	5.66 ^(+2.50)	5.49 ^(+0.93)	4.95 ^(+2.40)	5.73 ^(+2.21)
Qwen2.5VL-7B	2.88	4.20	3.29	4.02	3.98	2.24	4.77	2.93	3.54
+AnyCap-2B	3.39 ^(+0.51)	5.97 ^(+1.77)	6.23 ^(+2.94)	4.91 ^(+0.89)	5.23 ^(+1.25)	6.30 ^(+4.06)	5.51 ^(+0.74)	4.48 ^(+1.55)	5.25 ^(+1.71)
+AnyCap-8B	4.03 ^(+1.15)	6.15 ^(+1.95)	6.58 ^(+3.29)	4.81 ^(+0.79)	6.24 ^(+2.26)	6.04 ^(+3.80)	5.74 ^(+0.97)	4.82 ^(+1.89)	5.55 ^(+2.01)
InternVL3.5-8B	3.14	4.53	5.05	4.40	3.77	2.93	4.84	3.17	3.98
+AnyCap-2B	4.20 ^(+1.06)	5.88 ^(+1.35)	6.40 ^(+1.35)	5.33 ^(+0.93)	4.78 ^(+1.01)	5.58 ^(+2.65)	5.08 ^(+0.24)	4.72 ^(+1.55)	5.25 ^(+1.27)
InternVL3.5-14B	3.23	4.63	5.16	4.54	3.81	3.03	4.93	3.18	4.06
+AnyCap-2B	3.37 ^(+0.14)	5.44 ^(+0.81)	6.36 ^(+1.20)	4.73 ^(+0.19)	3.82 ^(+0.01)	4.83 ^(+1.80)	4.96 ^(+0.03)	4.42 ^(+1.24)	4.74 ^(+0.68)
InternVL3.5-38B	3.31	4.77	5.01	4.65	3.78	4.53	5.11	3.40	4.32
+AnyCap-2B	3.59 ^(+0.28)	5.52 ^(+0.75)	6.04 ^(+1.03)	4.70 ^(+0.05)	3.38 ^(-0.40)	4.84 ^(+0.31)	4.80 ^(-0.31)	4.18 ^(+0.78)	4.63 ^(+0.31)
Model	Brf.	Det.	Sty.	Poe.	Nar.	Avg.			
Proprietary Models									
GPT-4o	1.47	1.52	2.77	2.52	2.48	2.15			
+AnyCap-2B	1.97 ^(+0.50)	1.78 ^(+0.26)	2.52 ^(-0.25)	2.60 ^(+0.08)	2.65 ^(+0.17)	2.30 ^(+0.15)			
+AnyCap-8B	1.95 ^(+0.48)	1.82 ^(+0.30)	2.58 ^(-0.19)	2.50 ^(-0.02)	2.77 ^(+0.29)	2.32 ^(+0.17)			
Open-source Models									
InternVL2.5-8B	1.39	1.77	2.23	1.91	2.34	1.93			
+AnyCap-2B	1.84 ^(+0.45)	1.80 ^(+0.03)	2.45 ^(+0.22)	2.48 ^(+0.57)	2.39 ^(+0.05)	2.19 ^(+0.26)			
+AnyCap-8B	2.00 ^(+0.61)	1.88 ^(+0.11)	2.50 ^(+0.27)	2.20 ^(+0.29)	2.61 ^(+0.27)	2.24 ^(+0.31)			
Qwen2.5VL-7B	1.63	1.65	2.55	2.40	2.29	2.10			
+AnyCap-2B	1.66 ^(+0.03)	1.82 ^(+0.17)	2.48 ^(-0.07)	2.30 ^(-0.10)	2.55 ^(+0.26)	2.16 ^(+0.06)			
+AnyCap-8B	1.92 ^(+0.29)	1.98 ^(+0.33)	2.65 ^(+0.10)	2.45 ^(+0.05)	2.55 ^(+0.26)	2.31 ^(+0.21)			
InternVL3.5-8B	1.43	1.80	2.32	2.12	2.41	2.02			
+AnyCap-2B	1.82 ^(+0.39)	1.86 ^(+0.06)	2.37 ^(+0.05)	2.31 ^(+0.19)	2.44 ^(+0.03)	2.16 ^(+0.14)			
InternVL3.5-14B	1.45	1.86	2.37	2.21	2.48	2.07			
+AnyCap-2B	1.81 ^(+0.36)	2.16 ^(+0.30)	2.33 ^(-0.04)	2.33 ^(+0.12)	2.54 ^(+0.06)	2.23 ^(+0.16)			
InternVL3.5-38B	1.45	2.05	2.40	2.28	2.52	2.14			
+AnyCap-2B	1.93 ^(+0.48)	2.20 ^(+0.15)	2.36 ^(-0.04)	2.55 ^(+0.27)	2.57 ^(+0.05)	2.32 ^(+0.18)			

Table 26: Performance of additional models on the image modality of AnyCapEval.

Model	IPos.↑	IApp.↑	Ins.↑	Per.↑	Avg.↑	Brf.↑	Det.↑	Thm.↑	Poe.↑	Nar.↑	Avg.↑
LLaVA-7B	1.50	2.99	3.85	2.63	2.74	1.27	0.98	1.21	0.71	1.67	1.19
+AnyCap-8B	3.26	4.72	5.23	4.56	4.44	2.27	1.98	2.56	2.90	2.74	2.45
MiniCPM-o	1.96	3.80	4.90	3.07	3.43	2.11	2.06	2.15	1.74	2.54	2.12
+AnyCap-8B	2.31	4.40	5.00	2.98	3.67	2.36	2.14	2.73	2.33	2.74	2.44
YiVL-34B	1.38	1.85	1.74	1.73	1.67	0.91	0.59	0.85	0.55	1.04	0.81
+AnyCap-8B	3.54	4.69	4.93	4.66	4.45	1.98	2.04	2.59	2.65	2.59	2.30
InternVL2.5-38B	1.69	3.79	4.92	3.01	3.35	2.15	2.02	2.17	2.06	2.64	2.21
+AnyCap-8B	3.31	4.90	5.00	4.29	4.38	2.38	2.15	2.72	2.63	2.82	2.54

Table 27: Performance of additional models on the video modality of AnyCapEval. Top: Content metrics; Bottom: Style metrics.

Model	Content								
	IPos.↑	IApp.↑	Act.↑	Ins.↑	Pes.↑	Mov.↑	Bkg.↑	Evt.↑	Avg.↑
Tarsier2-7B	2.50	3.39	3.80	3.95	2.49	1.10	4.26	4.18	2.83
+AnyCap-8B	4.85	7.79	7.96	5.94	5.77	5.48	6.10	5.57	6.18
LLaVA-Video	2.77	5.89	2.84	3.67	3.10	1.34	6.15	2.65	3.55
+AnyCap-8B	4.37	7.08	6.20	5.68	5.79	5.35	6.62	4.35	5.68
MiniCPM-o	2.68	3.59	3.31	4.21	4.16	1.76	4.62	1.76	3.26
+AnyCap-8B	4.65	5.70	6.05	5.20	5.48	5.06	6.01	5.06	5.40

Model	Style					
	Brf.↑	Det.↑	Thm.↑	Poe.↑	Nar.↑	Avg.↑
Tarsier2-7B	0.63	2.08	2.12	0.42	2.45	1.51
+AnyCap-8B	1.87	1.90	2.65	2.38	2.55	2.26
LLaVA-Video	1.61	1.19	3.36	2.01	2.70	2.13
+AnyCap-8B	1.97	2.11	2.74	0.99	1.32	2.23
MiniCPM-o	1.29	1.40	2.30	1.73	2.48	1.84
+AnyCap-8B	1.97	2.11	2.74	0.99	1.32	2.23

Table 28: Video generation evaluation. Metrics: Visual Quality (VQ), Temporal Consistency (TC), Dynamic Degree (DD), Text–Video Alignment (TVA), and Factual Consistency (FC).

Model	VQ	TC	DD	TVA	FC	Avg
Wan2.1-T2V-14B	2.84	2.78	2.66	2.76	2.77	2.74
+AnyCap-8B	3.19 (+0.35)	3.06 (+0.28)	3.18 (+0.52)	3.07 (+0.31)	3.00 (+0.23)	3.10 (+0.36)
HunyuanVideo	2.98	2.89	2.94	2.90	2.83	2.91
+AnyCap-8B	3.61 (+0.63)	3.48 (+0.59)	3.62 (+0.68)	3.34 (+0.44)	3.46 (+0.63)	3.50 (+0.59)

Table 29: Human validation of AnyCapEval: ranking consistency with human judgments and inter-annotator agreement.

Dimension	Spearman ρ ↑	Kendall τ ↑	IAA
Content	0.93	0.86	0.88
Style	0.88	0.82	0.80

Table 30: Judge robustness: replacing GPT-4o with QwenVL3-8B on the image modality. Both absolute scores and relative rankings remain highly consistent.

Base Model	Content ↑		Style ↑	
	GPT-4o	QwenVL3	GPT-4o	QwenVL3
InternVL2.5-8B	3.04	3.05	2.12	2.15
+AnyCap	4.13	4.13	2.36	2.37
Qwen2.5-VL-7B	3.31	3.31	2.22	2.23
+AnyCap	4.36	4.35	2.50	2.50

Table 31: Robustness to judge-prompt paraphrasing via synonym substitutions. Rankings are stable under both mild and moderate perturbations.

Perturbation	Content	Style	Spearman ρ ↑
Original prompt	2.89	2.26	1.00
Synonym (mild)	2.90	2.22	1.00
Synonym (moderate)	2.89	2.25	1.00