Seeking Universal Shot Language Understanding Solutions

Anonymous CVPR submission

Paper ID 4

Abstract

Shot language understanding (SLU) is crucial for interpreting narrative, emotion, and aesthetic style in filmmaking. Although vision-language models (VLMs) demonstrate strong general capabilities, they struggle with the complex, multi-dimensional nature of cinematography, primarily due to a severe lack of high-quality training data. To bridge this gap, we introduce SLU-SUITE, a comprehensive dataset comprising 490K human-annotated image and video QA pairs that cover 33 distinct tasks across six cinematic dimensions. Leveraging SLU-SUITE, we originally identify that the core bottleneck of VLMs for SLU is semantic alignment with expert boundaries and terminology, rather than raw visual perception. Guided by this insight, we propose a balanced data scheduling and parameter-efficient strategy to train a universal SLU model, UniShot-8B. Comprehensive evaluations, including in-domain and out-of-domain settings, demonstrate the superiority of UniShot-8B.

1. Introduction

High-quality filmmaking relies not only on the script, but also on *shot language* [2, 8, 12] to convey narrative meaning, emotion, and aesthetic style. Shot language understanding (SLU) is highly complex: (1) it spans multiple aspects of cinematography, such as composition, shot scale, camera motion, lighting, colors, and inter-shot connections; and (2) even within a single aspect, different experts often adopt different taxonomies (e.g., different composition rules or different levels of granularity for camera motion).

Vision-language models (VLMs) show strong performance on general visual understanding. However, recent studies [4, 7, 8, 10, 11] reveal significant discrepancies between VLM judgments and film experts on SLU tasks. Consequently, there is a pressing need to develop universal models for SLU that can align well with film experts, handle diverse SLU tasks, and generalize to unseen scenarios.

However, a major obstacle is the lack of large-scale, high-quality training data. As shown in Table 1, existing human-annotated SLU datasets [7–10] usually cover only a small subset of tasks and are limited in scale. Meanwhile, current VLM-labeled datasets[6, 12], such as using GPT-4V, exhibits underlying misalignment issues. Thus, we introduce **SLU-SUITE**, a comprehensive training and evaluation suite for general SLU. SLU-SUITE organizes data into six film-grounded dimensions [2–4] (*composition*, *coverage*, *viewpoint*, *motion*, *lighting*, and *cuts*) and covers **33** detailed tasks with **490K human-labeled QA pairs**.

Building upon SLU-SUITE, we further study **VLM fine-tuning strategies for building universal SLU models**. We propose a simple data scheduling strategy to balance learning across different dimensions, together with a parameter-efficient adaptation strategy, i.e., finetuning the language model only. Specifically, we empirically demonstrate that the SLU performance of VLMs is mainly bottlenecked by *semantic alignment* rather than visual perception.

Based on these insights, we efficiently fine-tune Qwen3-VL-8B to obtain our universal SLU model, **UniShot-8B**. Through **evaluations on 23 SLU tasks**, UniShot-8B shows significant superiority: (1) in **in-domain (ID)** evaluation, one single UniShot-8B surpasses 12 task-specific fine-tuned VLMs in 9/12 cases; and (2) in **out-of-domain (OOD)** evaluation, UniShot-8B outperforms Gemini-3.0-Flash and Pro by an average 5%. accuracy improvement.

In short, our contributions are four-fold: the construction of the comprehensive dataset suite, the study of precise VLM fine-tuning strategies, the training of a universal SLU model, and comprehensive ID and OOD evaluations.

2. SLU-SUITE: Training and Evaluation Suite for General Shot Language Understanding

2.1. The Datasets

To mitigate the scarcity of data in shot language understanding (SLU), we build **SLU-SUITE**, a unified suite designed for *general* SLU training and evaluation. Compared with existing human-labeled cinematography datasets (Table 1), SLU-SUITE covers a broader set of tasks and explicitly supports both in-domain (ID) and out-of-domain (OOD) evaluation. As detailed in Table 2, SLU-SUITE spans six high-level dimensions, includes **33 detailed task**, and contains ∼**490K human-labeled QA pairs (including 53,607**

Human-Labeled Dataset	Scale	Training Supp.	OOD Eval.
ShotBench [8]	8 tasks ∼70k QA	✓	×
CameraBench [7]	5 tasks ∼150k labels	✓	×
CineTechBench [10]	7 tasks ∼0.7k QA	×	N/A
MovieCuts [9]	1 task ∼110k clips	✓	×
SLU-SUITE (Ours)	**33 tasks ∼490k QA**	✓	✓

Table 1. **Comparison with representative human-labeled cinematography datasets. Scale** reports the number of tasks and the approximate data size. **Training Supp.** indicates whether official splits support fine-tuning/training. **OOD Eval.** indicates whether the benchmark supports controlled out-of-domain evaluation by design (e.g., unseen sources/taxonomies/tasks). CineTechBench is evaluation-oriented without training partitions, thus OOD is marked as N/A. **Our SLU-SUITE** covers a broader set of tasks, provides with large-scale human-labeled data, and explicitly supports both in-domain (ID) and out-of-domain (OOD) evaluation to support general shot language understanding research, especially model finetuning/training.

image QA pairs and 433,718 video QA pairs).

Organizing 33 Tasks into 6 Film-Grounded Dimensions. Shot language understanding data are highly heterogeneous: different sources use different label spaces (e.g., coarse vs. fine-grained shot scale), some tasks are multi-label (option combinations), and some tasks are binary or free-form (caption). Training and evaluating on a flat list of tasks makes it hard to control coverage, prevent taxonomy confusion, and analyze transfer.

We therefore organize SLU-SUITE into **six high-level dimensions** that follow standard cinematography controls used in film analysis and on-set communication (e.g., framing/composition, scale and lensing, viewpoint, camera movement, lighting/color, and editing/cuts) and are consistent with structured filmmaking taxonomies [2–4].

As summarized in Table 2, each dimension contains multiple *task variants*:

- *Composition* (**3** tasks) captures how subjects and visual mass are arranged in the frame, including layout templates and weight/placement cues (*LayoutTemplate_**, *WeightPlacement*).
- *Coverage* (**9** tasks) models what a shot covers and how coverage is constructed, including shot scale taxonomies (*Scale_**), staging-based coverage patterns (*Staging*), mixed staging–scale labels (*StagingScaleMix*), and lens-related cues (*FocalLength*).
- *Viewpoint* (**6** tasks) describes camera placement relative to the subject and scene, including angle taxonomies (*Angle_**) and camera height (*Height*, *Height_Cartoon*).
- *Motion* (**10** tasks) covers both movement taxonomies and movement attributes: coarse vs. compound movements (*Move_**), motion description (*Move_Caption_**), and attribute-focused VQA such as speed, shaking, and complexity (*VQA_**).
- *Lighting* (**4** tasks) captures illumination and color style, including source/condition, lighting style/attributes, and palette cues (*SourceCondition*, *Style*, *Attribute*, *ColorPalette*).
- *Cuts* (**1** task) focuses on editing transitions within a shot-level context (*InerShotCutType*).

Note that, we consider semantically similar tasks as different detailed tasks when their option sets differs ((e.g., *Scale_Basic* vs. *Scale_Finegrained*).

Constructing via 11 Sources and 4 Stages. We build SLU-SUITE through a multi-source pipeline. We first identify 11 dataset sources that satisfy two requirements: (i) annotations are produced by humans (or verified by humans), rather than generated by LLMs; (ii) the sources jointly cover the six dimensions above, and each dimension is supported by at least two sub-tasks or has sufficient data volume (e.g., $\geq$50K). For each source, we collect the corresponding media and labels, and remove samples with missing files, broken links, corrupted content, or decoding failures. We then filter sensitive content using *Gemini-3.0-Flash*. Finally, we convert all sources into a unified QA format, including both classification-style QA and captioning-style QA. For classification-style QA, we provide explicit candidate options in the question and allow multi-label answers.

2.2. Evaluation Protocols

Mitigating media-level leak. Due to curating from multiple sources, we perform de-duplication across images and videos to ensure that evaluation media do not appear in the training set. After de-duplication, we obtain **about 53k unique images and 154k unique video clips**.

ID and OOD evaluation. We evaluate generalization in two settings: *ID*, where test tasks are included in training, and *OOD*, where test tasks are excluded from training.

For OOD evaluation, we hold out **11** classification-style QA tasks from two existing multi-task SLU datasets, *CineTechBench* [10] and *CameraBench* [7], and use them only for testing.

For the remaining **22** tasks, we select **12** tasks with more than 3K samples and split each of them into **80%** training and **20%** ID evaluation. We add all samples from the other **10** tasks to the training set. To further avoid cross-source leakage, we remove any OOD test sample whose media appears in the training set.

To keep it simple, we use captioning-style QA only for training, while both ID and OOD tests use classification-style QA with explicit answer options. Finally, we use around **410K** QA pairs for training and around **50K** QA pairs for evaluation.

High-Level Dimension	Detailed Tasks	Sources Num.	Samples Num.
Composition	**3** tasks: *CompositionRule*; *CompositionPattern*; *VisualWeightPlacement*	3	4,841
Coverage	**9** tasks: *Scale_Basic*;*Scale_Classic*;*Scale_Extended*; *Scale_Finegrained*; *Scale_Cartoon*; *Scale_Historical*; *Staging*; *StagingScaleMix*; *FocalLength*	7	51,693
Viewpoint	**6** tasks: *Angle_Basic*; *Angle_Extended*; *Angle_Finegrained*; *Angle_Cartoon*; *Height*; *Height_Cartoon*	4	37,742
Motion	**10** tasks: *Move_Coarse*; *Move_Compound_A*; *Move_Compound_B*; *Move_Captioning*; *VQA_Complexity*; *VQA_Movement*; *VQA_Shaking*; *VQA_Speed*; *VQA_Presence*; *VQA_MixedType*	5	282,297
Lighting	**4** tasks: *SourceCondition*; *Style*; *Attribute*; *ColorPalette*	2	925
Cuts	**1** task: *InterShotCutType*	1	109,827
Total	**33** tasks	**11** sources	**487,325** QA pairs

Table 2. **SLU-SUITE summary by high-level dimension.** Each row groups task variants under one dimension. **Detailed Tasks** lists task variants in that dimension (count in bold). We define a task variant by a specific option set (label space/taxonomy); thus, semantically similar tasks from different sources are treated as different variants when their option sets differ. **Sources Num.** denotes the number of contributing datasets, and **Samples Num.** denotes the number of labeled QA pairs. Overall, SLU-SUITE includes 33 task variants from 11 datasets, totaling 487,325 human-labeled QA pairs and covering 53,569 unique images and 154,451 unique videos after de-duplication.

3. Fine-Tuning Strategies for Universal SLU

Universal shot language understanding (SLU) requires learning many task variants with different taxonomies and formats, while keeping the underlying vision-language model (VLM) generalizable. We focus on two practical challenges: (i) **data strategy**—SLU-SUITE is highly imbalanced across dimensions and task variants (Table 2), so naive sampling can over-train a few large groups and hurt balanced performance; and (ii) **parameter strategy**—although modern VLMs are strong on generic vision-language tasks, recent benchmarks suggest they remain weak on SLU, where success often depends on expert terminology and taxonomy boundaries. Blind full-parameter fine-tuning is expensive and may also degrade the model's general capabilities, which can further reduce generalization. We therefore design a simple data schedule to balance training, and study where parameter-efficient adaptation should be applied to improve SLU most effectively.

Data strategy: balancing performance across dimensions. Sampling examples in proportion to dataset size is dominated by large dimensions (e.g., motion and cuts), which weakens balanced SLU competence. Leveraging the hierarchical organization of SLU-SUITE, we use a **dimension-balanced sampler**. For each mini-batch, we (1) sample one of the six high-level dimensions *uniformly*, and then (2) sample a QA instance from that dimension by first choosing a task variant at random and then choosing an instance from that variant. This schedule prevents large task groups from overwhelming training and encourages the model to revisit all dimensions throughout training.

Parameter strategy: precisely identifying the bottleneck of VLM. Given the same data strategy above, we study which component limits SLU performance in a typical VLM pipeline: the **vision encoder** (visual perception), the **multimodal projector** (vision–language connector), or the **language model** (semantic decision making). We conduct a controlled **adapter placement** study on Qwen3-VL-8B [1] using *Low-Rank Adaptation (LoRA)* [5]. Specifically, we compare five settings by enabling LoRA in: (i) none (no adaptation), (ii) vision encoder only, (iii) projector only, and (iv) language model only. (v) all layers. We control variables by using the *same training data*, *the same LoRA rank*, and *the same batch size and optimizer*, and only changing the trainable module.

Table 3. **SLU accuracy with different LoRA adapter placements in Qwen3-VL-8B.** All settings use the same data strategy and hyperparameters; only the adapter target differs.

	None	Vision Encoder	Projector	Language Model	All
Overall Acc. (↑)	0.467	0.567	0.571	<u>0.699</u>	**0.702**

As shown in Table 3, updating the **language model only** yields the largest gain, almost the same performance with updating all layers, but with significantly less parameters. This result suggests that VLMs' SLU performance is largely bottlenecked by *semantic alignment*—i.e., aligning with film experts' decision boundaries and taxonomy-specific terms—rather than capturing visual signals. Based on this finding, we adopt **language-model-only LoRA** as our default parameter-efficient fine-tuning strategy.

Implementation details of our SFT solution. Based on these, we fine-tune Qwen3-VL-8B with supervised fine-tuning (next-token prediction loss) for 3 epochs. We train on 2×A100 80GB GPUs, with batch size 4 per GPU and LoRA rank 32, within 2 days. We denote the resulting universal SLU model as **UniShot-8B**.

4. Experiments

We study two questions: (1) **ID performance:** can our universal multi-task model match or exceed other general or even task-specific solutions? (2) **OOD generalization:** does the model generalize well to unseen tasks?

We compare our **UniShot-8B** with: **Five Universal (one-for-all) Solutions.** We evaluate: (i) Qwen3-VL-8B [1] as an open-source VLM baseline; (ii) *Gemini-2.5-Flash, Gemini-3.0-Flash, Gemini-3.0-Pro* as powerful commercial VLMs; (iii) *ShotVL-7B [8]* as a SOTA multi-task SLU model via finetuning Qwen2.5-VL-7B; and **Task-specific**

Table 4. **In-domain (ID) results on 12 tasks.** We compare *Universal* (one model for all tasks) against *Specific* (one model per task). **Count #1** reports how many task variants each method ranks first.

	Universal (one-for-all)						Specific (one-for-one)
12 In-Domain Tasks	Qwen3-VL-8B	Gemini-2.5-Flash	Gemini-3.0-Flash	Gemini-3.0-Pro	ShotVL-7B	**UniShot-8B (Ours)**	12 task-specific-SFT Qwen3-VL-8B
CompositioRule	0.020	0.005	0.040	0.040	0.775	**0.791**	0.784
ScaleExtended	0.145	0.165	0.125	0.170	0.115	0.290	**0.320**
ScaleHistorical	0.670	0.770	0.850	0.825	0.740	**0.900**	0.855
ScaleCartoon	0.715	0.675	0.650	0.660	0.460	**0.940**	0.875
ScaleBasic	0.650	0.465	0.595	0.555	0.425	**0.875**	0.865
Staging	0.430	0.165	0.280	0.295	0.290	**0.785**	0.685
AngleBasic	0.330	0.420	0.440	0.485	0.245	0.940	**0.960**
AngleCartoon	0.350	0.515	0.450	0.545	0.260	0.690	**0.710**
Height	0.630	0.690	0.705	0.670	0.650	**0.905**	0.875
HeightCartoon	0.540	0.385	0.555	0.580	0.600	**0.740**	0.720
MoveCoarse	0.745	0.650	0.710	0.585	0.600	**0.850**	0.840
InterShotCutType	0.230	0.215	0.295	0.270	0.210	**0.400**	0.390
Avg. Acc. (↑)	0.455	0.427	0.475	0.473	0.448	**0.759**	0.740
Count #1 (↑)	0/12	0/12	0/12	0/12	0/12	**9/12**	3/12

Table 5. **Out-of-domain (OOD) results on 11 unseen task variants** (Universal solutions only). OOD tasks use unseen task forms and/or new taxonomies from different sources.

11 Out-of-Domain Tasks	Qwen3-VL-8B	Gemini-2.5-Flash	Gemini-3.0-Flash	Gemini-3.0-Pro	ShotVL-7B	**UniShot-8B (Ours)**
CompositionPattern	0.083	0.208	0.158	0.525	0.133	**0.558**
ScaleClassic	0.371	0.486	**0.600**	0.436	0.507	0.414
FocalLength	0.217	0.333	**0.467**	0.433	0.300	0.367
AngleExtended	0.650	0.575	**0.825**	0.792	0.608	0.608
MoveCompound	0.815	0.750	0.810	0.585	0.805	**0.880**
VQAMovement	0.366	0.452	0.570	0.441	0.301	**0.634**
VQAComplexity	0.535	0.555	0.570	0.355	0.490	**0.800**
VQAShaking	0.640	0.615	0.565	0.520	0.615	**0.720**
VQASpeed	0.510	0.680	0.605	0.645	**0.730**	0.675
LightingStyle	0.582	0.591	0.600	0.691	0.582	**0.700**
ColorPalette	0.483	0.600	0.617	0.600	0.550	**0.633**
Avg. Acc. (↑)	0.478	0.531	0.581	0.547	0.511	**0.636**
Count #1 (↑)	0/11	0/11	3/11	0/11	1/11	**7/11**

(one-for-one) Solutions. We also consider *12 task-specific-SFT Qwen3-VL-8B models* in ID evaluations, each fine-tuned on a single training task corresponding to the same test task.

All evaluations use classification-style QA. Accuracy is reported over three runs with decoding temperature as 0. For multi-label questions, we use strict exact-match.

ID Results: one UniShot-8B beats 12 task-specific fine-tuning models. Table 4 summarizes results on 12 ID task variants. UniShot-8B achieves the best average accuracy (**0.759**) and ranks first on **9/12** tasks, substantially improving over the base Qwen3-VL-8B and the general-purpose VLMs. Notably, UniShot-8B also slightly exceeds the task-specific upper bound on average (0.759 vs. 0.740), suggesting that *multi-task training can provide positive transfer across SLU tasks, where learning one task can strengthen others compared to training them independently.* Overall, the ID results support our goal of a universal SLU model that covers many tasks without sacrificing accuracy.

OOD Results: UniShot-8B generalizes well to unseen SLU tasks. Table 5 reports performance on 11 OOD task variants, which include unseen task forms and/or new taxonomies from different sources. UniShot-8B achieves the best average accuracy (**0.636**) and ranks first on **7/11** tasks, outperforming the open-source baselines and the stronger commercial model Gemini-3.0-Flash on average. *The OOD results demonstrate that large-scale multi-task training enables UniShot-8B with strong generalization ability, thereby broadly supporting unseen tasks.*

Other Insights. (I) We also observe that **composition** remains a clear weakness for general-purpose VLMs: most general VLM baselines fail on the two composition tasks, with accuracies staying below 20%. (II) In addition, ShotVL-7B is limited by its narrower SLL task coverage: although its overall SLU performance is comparable to that of Gemini-2.5-Flash, it shows notable generalization failures on several uncovered tasks and even underperforms the similarly sized Qwen3-VL-8B. These results further motivate the need for a comprehensive and large-scale suite like our SLU-SUITE to support universal SLU models building.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 3

[2] David Bordwell, Kristin Thompson, and Jeff Smith. *Film art: An introduction*. McGraw-Hill New York, 2008. 1, 2

[3] Blain Brown. *Cinematography: theory and practice: image making for cinematographers and directors*. Routledge, 2016.

[4] Agneet Chatterjee, Rahim Entezari, Maksym Zhuravinskyi, Maksim Lapin, Reshinth Adithyan, Amit Raj, Chitta Baral, Yezhou Yang, and Varun Jampani. Stable cinemetrics: Structured taxonomy and evaluation for professional video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 1, 2

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022. 3

[6] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024. 1

[7] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Rushikesh Zawar, Xue Bai, et al. Towards understanding camera motions in any video. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 1, 2

[8] Hongbo Liu, Jingwen He, Yi Jinn, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Weichao Chen, Yu Qiao, et al. Shotbench: Expert-level cinematic understanding in vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 1, 2, 3

[9] Alejandro Pardo, Fabian Caba Heilbron, Juan León Alcázar, Ali Thabet, and Bernard Ghanem. Moviecuts: A new dataset and benchmark for cut type recognition. In *European Conference on Computer Vision*, pages 668–685. Springer, 2022. 2

[10] Xinran Wang, Songyu Xu, Shan Xiangxuan, Yuxuan Zhang, Muxi Diao, Xueyan Duan, Kongming Liang, Zhanyu Ma, et al. Cinetechbench: A benchmark for cinematographic technique understanding and generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 1, 2

[11] Hang Wu, Yujun Cai, Haonan Ge, Hongkai Chen, Ming-Hsuan Yang, and Yiwei Wang. Refineshot: Rethinking cinematography understanding with foundational skill evaluation. *arXiv preprint arXiv:2510.02423*, 2025. 1

[12] Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28984–28994, 2025. 1