Generalizing Sports Feedback Generation by Watching Competitions and Reading Books: A Rock Climbing Case Study

Arushi Rai and Adriana Kovashka

Motivation

- Coaches help guide athletes to improve their technique through feedback
- This feedback is specific to their performance and are generally actionable
- Training feedback generation models requires expensive-to-collect feedback from the target sport despite the prevalence of video-LLMs.

Problem 1: Feedback generation models fail to generalize to sports out of training domain.




The climber is relying too much on upper body strength and not utilizing their lower body enough, leading to poor positioning and energy expenditure.

🟢 **In-Domain** 🤖

The shooter's follow-through and wrist snap are good, but the left hand needs to be more involved in the shot.

🔴 **Out-Of-Domain** 🤖

Lost ability to recognize sport and hallucinates

Problem 2: Semantic evaluation metrics don't capture relevant feedback quality aspects

Generated	Reference
"The shooter is shooting the ball poorly"	"**The ball's trajectory is flat because the release point is too late. This is because the shoulders and hips are slow to rotate**, so **rotate the hips faster.**"

Generated feedback evaluated with reference misses aspects present in the reference data such as the precision of details about body parts and movement quality and actionability of feedback

Source Domain (Basketball, Soccer)

The player demonstrates good technique in juggling the ball, maintaining control and using the lower leg to strike the ball at the correct contact point.

The follow-through hand is directing the ball to the left side of the basket, and it needs to be adjusted to guide the ball to the middle of the targeted area.

The shooter's follow-through is perfect, with a great wrist flick and good lift on the right side.

The player's arms should be further out for better balance, especially at higher speeds.

Target Domain (Rock Climbing)

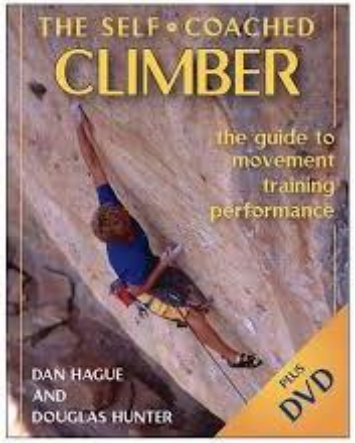
YouTube

Competition Videos

[00:35:31 → 00:35:44]
"So if I hang like this, uh, the first thing I would definitely do is to engage my shoulder to go up, uh... but I'm going to go back there going, yeah, that's..."

Refinement
+ Precise Localization

[00:35:33 → 00:35:35]
Engage shoulder to initiate upward movement



Coaching Textbooks

Annotated Paired Video-Feedback Data

Used for training

Video-LLM

Auxiliary Multi-Modal Web Data

Used for training

Source Domain (Basketball, Soccer)

The player demonstrates good technique in juggling the ball, maintaining control and using the lower leg to strike the ball at the correct contact point.

The follow-through hand is directing the ball to the left side of the basket, and it needs to be adjusted to guide the ball to the middle of the targeted area.

The shooter's follow-through is perfect, with a great wrist flick and good lift on the right side.

The player's arms should be further out for better balance, especially at higher speeds.

Target Domain (Rock Climbing)

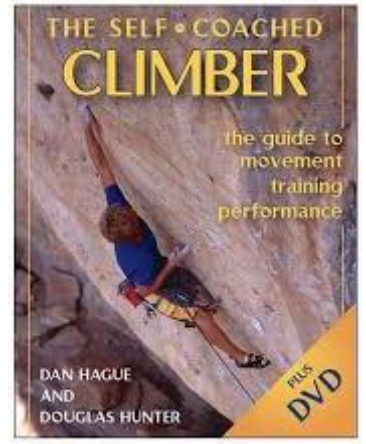
YouTube

Competition Videos

[00:35:31 → 00:35:44]
"So if I hang like this, uh, the first thing I would definitely do is to engage my shoulder to go up, uh... but I'm going to go back there going, yeah, that's..."

Refinement
+ Precise Localization

[00:35:33 → 00:35:35]
Engage shoulder to initiate upward movement



Coaching Textbooks

Annotated Paired Video-Feedback Data

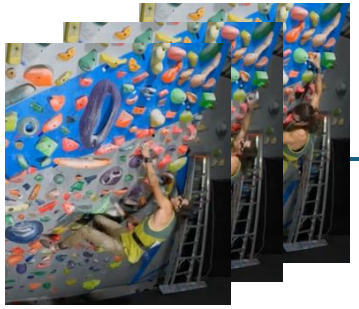
Used for training

Auxiliary Multi-Modal Web Data

Used for training

Video-LLM

Inference



Video-LLM

"The climber successfully made the next move by twisting their body and pivoting around their right shoulder, allowing them to hang on to the handhold."

Evaluation

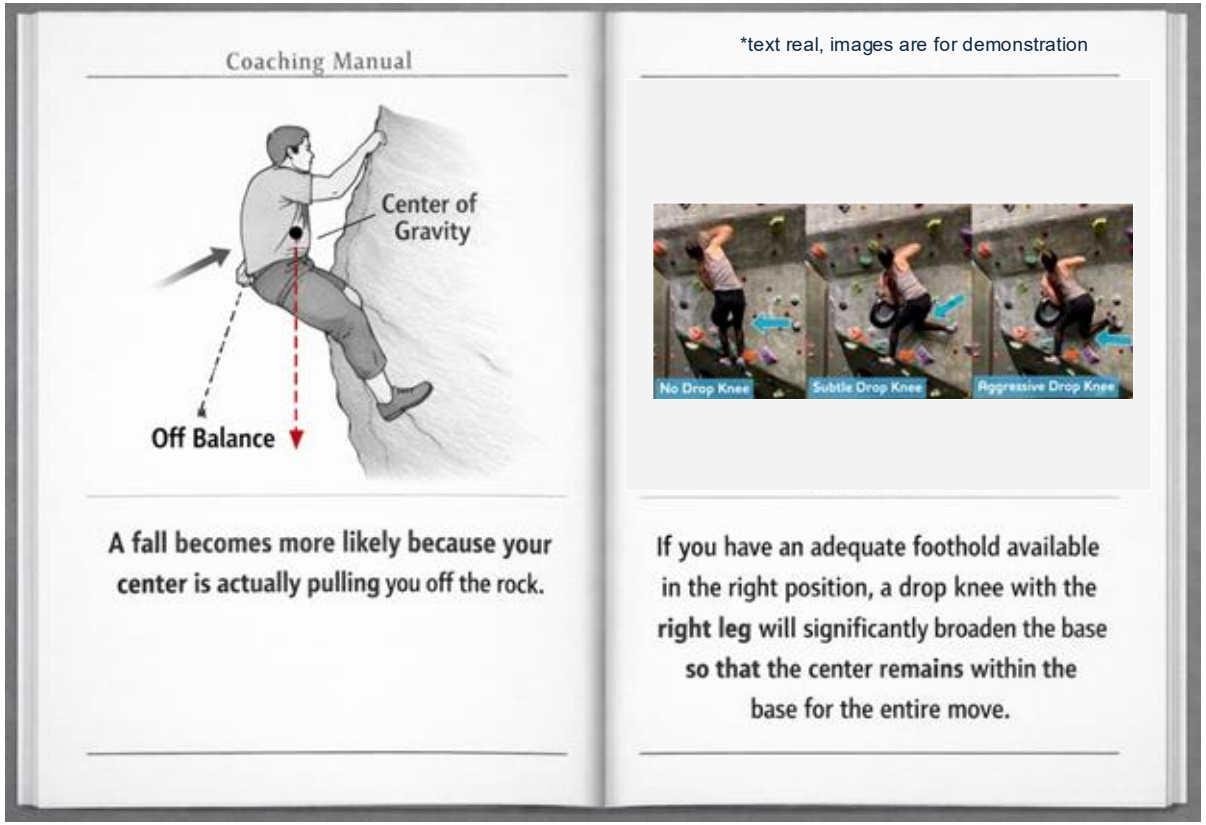
- Lexical & Semantic Similarity Metrics (BLEU-4, METEOR, ROUGE-L, BERTScore)
- Our Proposed LLM-based Metrics
  - **Specificity**
  - **Actionability**

Training data

Looks up to anticipate the next moves. **Timestamp (s): [28.0, 30.26]**

Climber is performing a crimp, making it appear effective. **Timestamp (s): [19.66, 23.66]**

Cleaner, more aligned commentary-video pairs



Examples of coaching text that illustrate cause and effect of body positioning.

Experiment Details

- Unified training on coaching text and source video-feedback data+competition video-commentary pairs through training on next-token prediction task.
- Model:
  - InternVideo2.5 [2]
- Data
  - Source video-feedback from ExpertAF [1] (18K), our competition video-commentary pairs (15K)
  - 2,806 test rock climbing samples from ExpertAF
- Metrics
  - BLEU-4, ROUGE-L, METEOR, BERTScore
  - Ours: Specificity and Actionability (introduced next)

[1] K. Ashutosh, T. Nagarajan, G. Pavlakos, K. Kitani, and K. Grauman,
"ExpertAF: Expert Actionable Feedback from Video," *CVPR 2025.*
[2] Y. Wang *et al.*, "InternVideo2.5: Advancing Vid Foundation Models for Multimodal Understanding," 2025.

Proposed evaluation metrics (Specificity)

- Specificity rates how precisely the feedback captures details about the learner's movement at the current moment.
  - Higher specificity includes detailed information about the movement (e.g., positioning, timing) and links it to quality indicators like smoothness, control, or efficiency.
  - 4-point scale

Level 1 (Least Specific)	Level 2 (Vague)	Level 3 (Slightly Specific)	Level 4 (Very Specific)
The shot could be improved.	The shooter is standing up straight.	Standing straight up limits explosiveness and lift.	Standing straight up limits explosiveness and lift because it prevents your lower body from fully loading the muscles needed for an explosive push-off.
The shot is poor.	Your arm was bent too much.	Your arm was bent too much causing the shot to look stiff.	Your guide arm was bent too much prior to lifting up to the release point, and caused the shot to look stiff.

Proposed evaluation metrics (Actionability)

- Actionability rates if a learner can directly apply the feedback to make a change. More actionable feedback gives clear corrective directions (e.g., what to move, how to adjust)
  - 3-point scale

Level 1 (Not Actionable)	Level 2 (Minimally Actionable)	Level 3 (Actionable)
That wasn't quite right.	Your stance is off-balance.	Widen your stance to be shoulder-length apart and keep your weight centered over your feet to maintain balance.
The climber could use a more efficient technique.	The climber is using a one-hand hold start, which is a good technique for beginners, but may not be the most efficient for experienced climbers.	For a more efficient climb, try switching from a one-hand hold start to a two-handed start and engage both your hands and core simultaneously so you can distribute your weight evenly.

Skipped-If the feedback is only positive reinforcement.

Results – Adding auxiliary data (Pt. 1)

	BLEU-4	METEOR	ROUGE-L	BERT
ID Fd.	3.31	20.94 ± 0.14	25.91 ± 0.13	37.3
Zero-Shot	1.75	15.08 ± 0.12	19.78 ± 0.04	30.3
OOD Fd.	1.30	11.45 ± 0.12	17.30 ± 0.11	25.4
Ours	**2.68**	**15.59 ± 0.14**	**24.01 ± 0.05**	**37.9**

Table 2. Out-of-distribution feedback generation evaluation. ID represents an upper bound using in-domain feedback data. Bold represents best result in column. Fd=Feedback. Ours is a combination of training on OOD feedback and our auxiliary sources.

Improvement across all text generation metrics compared to zero shot and training on only out-of-domain Feedback.

Close semantic similarity performance of the upper bound.

Results – Adding auxiliary data (Pt. 2)

Data Type	METEOR	ROUGE-L	BERT
Zero-Shot	15.08 ± 0.12	19.78 ± 0.04	30.3
Text	15.22 ± 0.06	19.74 ± 0.04	30.4
Com., Fd.	15.38 ± 0.10	23.39 ± 0.06	37.0
T., Com., Fd.	**15.59** ± 0.14	**24.01** ± 0.05	**37.9**

Table. Data source ablation. **Adding commentary data boosts performance over zero-shot and text-only fine-tuning.** Training with all data types performs best across all metrics.
Com=Commentary, Fd=Feedback.

Results on our proposed metrics

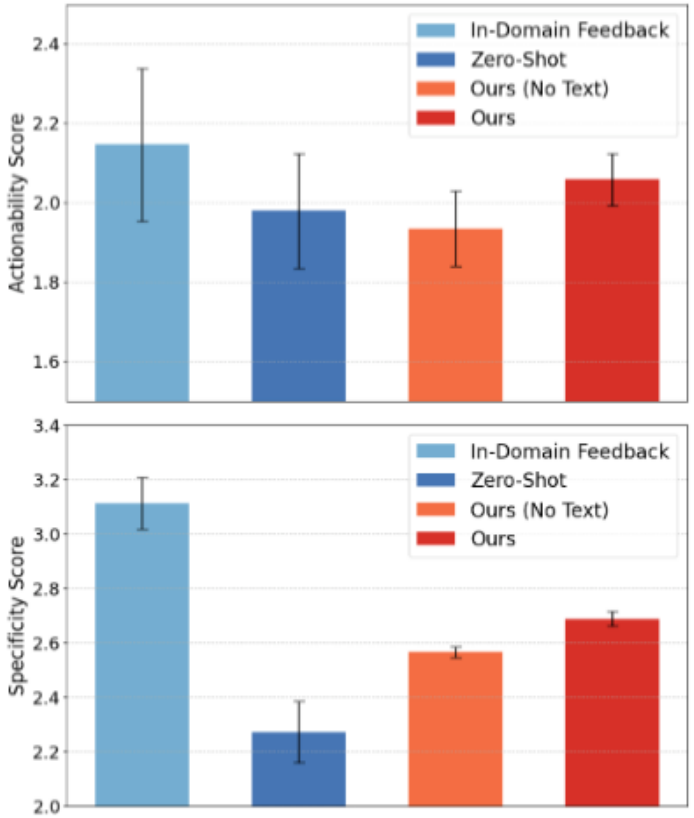


Figure 3. LLM-based evaluation of actionability (top) and specificity (bottom) as introduced in Sec. 5.2. The lines indicate max/min over using GPT-4o, Gemini 2.5, and DeepSeek Chat.

Inclusion of text-only data has significant improvements in actionability compared to other all metrics and steady improvements on specificity.

Conclusion

- Auxiliary freely-available web data (competition videos and coaching manuals) significantly improves target domain feedback generation performance without collecting in-domain feedback data.
- Proposed **specificity** and **actionability** metrics capture unique aspects of feedback quality that traditional NLP metrics miss.
- Together, our approach enables more meaningful and practical generation of sports feedback under limited annotations.