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

1. Commentary and video timing

Our proposed precise localization method aims to pinpoint when the refined commentary occurs based on word-level timestamps. This solves temporal misalignment caused by the coarseness of the original ASR timestamps, but there still can exist temporal misalignment between the narration and the corresponding video segment. However, this is less problematic for rock climbing since climbers often remain in position for extended periods, and commentary typically addresses actions that persist over time. We demonstrate this property with examples in Fig. 1. For sports beyond rock climbing that have short, fast-moving actions or events, finer temporal localization methods may be required (e.g., predicting the offset from the localized narration timestamp).

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]**

Figure 1. Examples of key-frames associated with refined commentary and timestamp window from precise localization step.

We also show an ablation in Fig. 2 with different windowing around the predicted timestamp. In particular, we match the narration to video content that comes *before* the predicted start and end times. This is to test whether video and narration are reasonably aligned, or if the narration captures actions that took place much earlier in the video. We observe that different windowing strategies around the predicted timestamps make a minimal difference and all perform better than the approach without precise localization described in Sec. 3 on most metrics.

Order	BLEU-4	METEOR	ROUGE-L
Sequential	**2.68** $\pm$ 0.02	**15.59** $\pm$ 0.14	**24.01** $\pm$ 0.05
Joint	2.63 $\pm$ 0.03	15.47 $\pm$ 0.04	23.87 $\pm$ 0.04

Table 1. Training order ablation. Sequential is training on text first, then on the video-text data. Joint is training on both data types together.

2. Does the order of training on auxiliary data affect generalization?

Next, we include ablations on the order of training on the auxiliary multimodal data and each stage of the data processing pipeline.

In Tab. 1, we compare two strategies for incorporating multiple data sources: sequential training and joint training. In the sequential setup, we first train on text-only data, followed by fine-tuning on video-text (commentary + OOD) data. In the joint setup, all auxiliary data sources and OOD feedback data are combined and trained on simultaneously. We find that sequential training slightly outperforms joint training across all metrics, suggesting that first learning the domain-relevant language prior to learning video-conditional generation may facilitate better generalization. However, the differences are relatively small, indicating that both strategies are viable for leveraging auxiliary data.

3. Does our pipeline to improve the quality of the competition data improve generalization ability?

We validate each design decision of our pipeline: training with OOD feedback data, refining the in-the-wild commentary data, and performing precise localization. To test the impact of skipping precise localization without decreasing the sampling rate or excessive memory usage, we produce 4-second clips by running a sliding window through the ASR segment. Each clip is paired with the refined commentary from its corresponding ASR segment.

In Tab. 2, we see that using the noisy in-the-wild commentary data alone yields worse performance compared to zero-shot performance in Tab. 1 in the main paper. Combin-

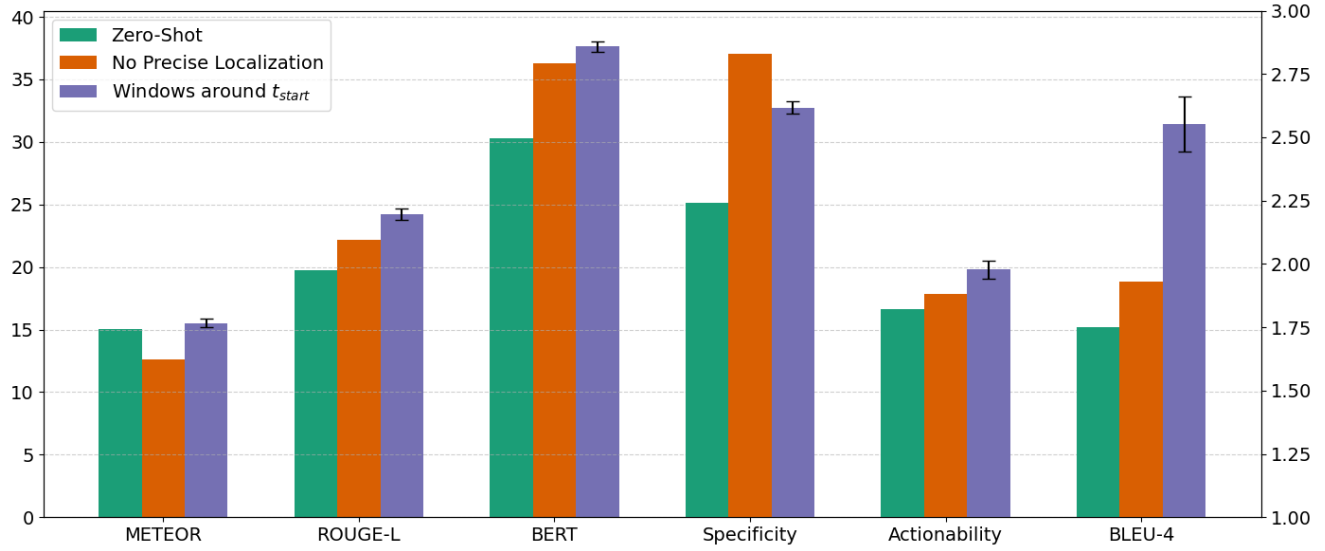

Figure 2. **Window ablation.** t_{start} and t_{end} are the start and end timestamps produced by the precise localization step. For *Windows around* t_{start}, the mean and confidence interval are computed over the performance of the following windowing strategies when used for training: $(t_{\text{start}}, t_{\text{end}})$, $(t_{\text{start}} - 3, t_{\text{start}} + 1)$, $(t_{\text{start}} - 4, t_{\text{start}})$, and $(t_{\text{start}} - 4, t_{\text{end}})$. These windows experiment with different ways of including actions that may have occurred prior to the narration. Observe that the confidence interval is very small, indicating comparable performance when the window is slightly shifted earlier, but consistent improvement over lack of precise localization.

OOD Fd.	R	PL	BLEU-4	METEOR	ROUGE-L	BERT
			1.06 ± 0.02	7.50 ± 0.06	13.65 ± 0.14	15.6
✓			1.62 ± 0.04	13.44 ± 0.05	19.20 ± 0.06	29.1
✓	✓		1.93 ± 0.01	12.60 ± 0.07	22.18 ± 0.10	36.3
✓	✓	✓	2.67 ± 0.02	15.38 ± 0.10	23.39 ± 0.06	37.0

Table 2. Ablation of each stage of our commentary data processing pipeline and effect of adding OOD feedback from the source domain. Significant improvements from using the commentary data are from refinement and precise localization. R=refinement. PL=precise localization.

ing with OOD feedback data improves performance, however, the most significant improvements come from refinement and the precise localization stages of the data processing pipeline.

4. LLM refinement

The supplementary file also provides all prompt templates and annotation references used in our study. These include prompt instructions provided to LLMs for ASR commentary refinement, precise localization of refined commentary, and rating of feedback specificity and actionability.

To refine noisy ASR outputs from competition videos into concise and anonymized commentary, we prompt an LLM to remove irrelevant details and retain key information related to pose, body movement, and quality of execution. The prompt below was used to guide this refinement process.

LLM Refinement Prompt

You are an ASR refiner for a rock climbing competition. You take ASR inputs which may be noisy and not concise and output concise commentary. Replace named entities with general terms (e.g., "person", "competition"). Focus on capturing the action, body parts, pose information, and quality of movement. If the input is unintelligible or contains only music or applause, return only: `[SKIP]`.

Input:

`ASR narration: {narration}`

Output:

Cleaned narration:

5. Precise localization of refined commentaries

To enable more precise timestamps of refined commentary in ASR transcripts, we prompt an LLM to localize each commentary segment to a short (1–4 second) time span

based on context provided by the word-level timestamps. See prompt below.

Precise localization of refined commentaries Prompt

You are an ASR refiner for rock climbing competition commentary. Your task is to match cleaned, anonymized commentary with corresponding timestamps in noisy ASR input. The ASR input consists of transcriptions with word-level timestamps. Your goal is to determine when each action in the cleaned commentary occurs in the ASR by finding the most relevant timestamps. Each action should be localized to a **1s–4s time span** based on matching words and phrases. If an exact match is unavailable, use the closest approximation. Do **not** return timestamps for unintelligible sections (music, applause, background noise). Ensure all timestamps are precise and correspond to the moment when the action is happening.
Format the output as a structured list where each commentary line is paired with its estimated time range (corresponding **1s–4s** range) in the ASR. Please be concise.
Example output format:

> ["commentary": "The climber hooks the toe on the right and pulls himself up.", "timestamp": (47.8, 49.66), "commentary": "He reaches for the next hold.", "timestamp": (50.5, 52.3)]

Return only the JSON list—no other explanations, markdown, or formatting characters.
ASR with word-level timestamps: {`whisper_transcript`}
Refined commentary: {`refined_commentary`}
Output:

6. Automatic evaluation prompts

We assess the specificity and actionability of generated feedback. This LLM-based rating helps us evaluate feedback quality more interpretably. See the prompts below.

Table 3 presents the full set of specificity examples (levels 1–4) shown to annotators. These show how feedback becomes more informative and elaborative as specificity increases. Similarly, Table 4 shows feedback examples at each level of the actionability scale (1–3). Both of these tables were used to train annotators and calibrate (via in-context learning) LLM scoring models.

Specificity Rating Prompt

Analyze the given generated feedback and provide only a numerical rating for the **generated feedback**, from **1 to 4**, where 1 means "not specific" and 4 means "very specific".
Definition: Feedback conveys details about current movement and corrective measures. Specificity refers to the precision of movement information, focusing on the present; actionability guides future adjustments.
Ratings Guide:

- **1 – Least Specific:** Very vague, offers little useful information.
- **2 – Vague:** Mentions either movement pattern details *or* quality descriptors (e.g., smoothness, stiffness), or just performance outcomes.
- **3 – Slightly Specific:** Connects movement details to quality indicators but lacks elaboration.
- **4 – Very Specific:** Precise movement and quality info with elaboration (e.g., when, why, or how).

Examples:

```
Rating '1':
  - "The shot could be improved."
Rating '2':
  - "The shooter is standing up
   ↪  straight"
Rating '3':
  - "Standing straight up limits
   ↪  explosiveness and lift"
Rating '4':
  - "Standing straight up limits
   ↪  explosiveness and lift because
   ↪  it prevents your lower body
   ↪  from fully loading the muscles
   ↪  needed for an explosive
   ↪  push-off."

Rating '1':
  - "The climber needs to have more
   ↪  confidence."
Rating '2':
  - "The climber hesitates before
   ↪  reaching for the higher hold"
Rating '3':
  - "The climber hesitates and takes
   ↪  a shorter step when reaching
   ↪  for the higher hold."
Rating '4':
  - "The climber hesitates and takes
   ↪  a shorter step when reaching
   ↪  for the higher hold, which
   ↪  limits the momentum needed to
   ↪  successfully grab it."

Rating '1':
  - "The player is dribbling poorly"
Rating '2':
```

```
  - "The contact with the ball is
   ↪  closer to the heel rather than
   ↪  through that inside curvature
   ↪  of the foot. "
Rating '3':
  - "The contact with the ball is
   ↪  closer to the heel rather than
   ↪  through that inside curvature
   ↪  of the foot which affects
   ↪  controllability"
Rating '4':
  - "The contact with the ball is
   ↪  closer to the heel...so we have
   ↪  less control over the direction
   ↪  of the pass."
```

Actionability Rating Prompt

Analyze the generated feedback and provide only a numerical rating for the **generated feedback**, from **1 to 3**, where 1 means "not actionable" and 3 means "actionable".

Definition: Actionability refers to the degree to which feedback can be implemented by the learner (e.g., specific corrective directions). This scale evaluates how directly feedback helps guide performance adjustments. Skip scoring if the feedback is purely positive reinforcement.

Scale:

- **Skipped** – If the feedback is only positive reinforcement.
- **1 – Not Actionable:** Vague or lacks any clear guidance the learner can act on.
- **2 – Minimally Actionable:** Identifies what to change, but not how to do it.
- **3 – Actionable:** Provides specific, clear directions to help the learner adjust.

Example Progressions:

```
Rating '1':
- "That wasn't quite right."
Rating '2':
- "Your stance is off-balance."
Rating '3':
- "Widen your stance to be
 ↪  shoulder-length apart and keep
 ↪  your weight centered over your
 ↪  feet to maintain balance."

Rating '1':
- "The climber could use a more
 ↪  efficient technique."
Rating '2':
- "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."
Rating '3':
- "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."

Rating '1':
- "Your form is poor."
Rating '2':
- "Your arm bent too much."
Rating '3':
- "Keep your arm straight until you
 ↪  initiate the follow-through."

Rating '1':
- "The player is dribbling poorly."
Rating '2':
- "The players dribble lacks control
 ↪  because their touches are
 ↪  inconsistent."
Rating '3':
- "Use smaller, more controlled
 ↪  touches on the ball and stay on
 ↪  the balls of your feet to
 ↪  maintain better control."

Rating '1':
- "The player's first touch was off."
Rating '2':
- "The first touch is slow and takes
 ↪  them in the wrong direction."
Rating '3':
- "The player's first touch is slow
 ↪  and takes them in the wrong
 ↪  direction, causing them to take
 ↪  an extra touch and lose time.
 ↪  They need to move their feet in
 ↪  the direction they want to go."

Rating '1':
- "The ball's trajectory was off."
Rating '2':
- "The balls trajectory is too flat."
Rating '3':
```

```
- "Release the ball slightly earlier
 ↪  and follow through higher to
 ↪  create a better arc on your
 ↪  shot."

Rating '1':
- "The climber is struggling."
Rating '2':
- "The climber should work on
 ↪  improving their grip strength and
 ↪  endurance through training."
Rating '3':
- "The climber could improve their
 ↪  grip strength and endurance by
 ↪  incorporating more finger
 ↪  exercises and grip strengthening
 ↪  exercises into their training
 ↪  routine."

Rating '2':
- "Could improve by keeping their
 ↪  feet closer together and using
 ↪  their hips to generate power."
Rating '3':
- "Improve by adjusting foot
 ↪  positioning and engaging the hips
 ↪  more."
```

7. Instructions for annotators

We include tables showing example feedback at each specificity and actionability level, which were provided to annotators.

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.
The player missed the shot.	The ball's trajectory is flat.	The ball's trajectory is flat because the release point is too late.	The ball's trajectory is flat because the release point is too late. This is because the shoulders and hips are slow to rotate.
The climber needs to have more confidence.	The climber hesitates before reaching for the higher hold.	The climber hesitates and takes a shorter step when reaching for the higher hold.	The climber hesitates and takes a shorter step when reaching for the higher hold, which limits the momentum needed to successfully grab it.
The woman is doing a good job climbing the wall.	The climber's movements are smooth and controlled.	The climber is executing a great job, with a smooth and controlled movement, especially in transitions between holds.	The climber is executing a great job, with a smooth and controlled movement due to excellent foot placement and efficient weight transfer, especially in transitions between holds.
The climber has good technique.	The climber maintains good control.	The climber has successfully placed their right foot on a ledge and released their left foot to add force which helps with control.	The climber has successfully placed their right foot on a ledge, applying sufficient pressure, and released their left foot to add force to the right foot, which will help them stay pulled into the wall leading to more control.

Table 3. Examples of feedback at different specificity levels (1–4).

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.
Your form is poor.	Your arm bent too much.	Keep your arm straight until you initiate the follow-through.
The player is dribbling poorly.	The player's dribble lacks control because their touches are inconsistent.	Use smaller, more controlled touches on the ball and stay on the balls of your feet to maintain better control.
The player's first touch was off.	The first touch is slow and takes them in the wrong direction.	The player's first touch is slow and takes them in the wrong direction, causing them to take an extra touch and lose time. They need to move their feet in the direction they want to go.
The ball's trajectory was off.	The ball's trajectory is too flat.	Release the ball slightly earlier and follow through higher to create a better arc on your shot.
The climber is struggling.	The climber should work on improving their grip strength and endurance through training.	The climber could improve their grip strength and endurance by incorporating more finger exercises and grip strengthening exercises into their training routine.
–	Could improve by keeping their feet closer together and using their hips to generate power.	Improve by adjusting foot positioning and engaging the hips more.

Table 4. Examples of feedback at different actionability levels (1–3).

8. LLM biases

LLMs are known to carry various biases. This can be problematic when using LLMs as an evaluator. To explore these biases within the context of using an LLM to automatically score specificity and actionability, we consider two types of bias: gender and length bias.

To assess gender bias, we rewrite a set of feedback with male and female pronouns and compute specificity and actionability scores for each. First, we randomly sampled 20 feedbacks from the ExpertAF dataset. For each of the 20 examples, we manually created a version that uses male pronouns and another that uses female pronouns. We then calculated the specificity and actionability scores using GPT4o. There were no differences in the actionability and specificity scores. This may indicate that biases from LLMs may occur in open-ended generation and more abstract ratings such as sentiment, in contrast, specificity and actionability judge the structure and the definitions are not dependent on gendered pronouns.

To assess length bias, we report the effect of adding a neutral phrase to increase the feedback length. We append a neutral phrase (i.e. "This was observed during practice") that should not impact specificity or actionability. We observe a 0.05 increase in specificity (2.95 to 3.00) and actionability (2.2 to 2.25) when the neutral text is appended to the feedback. This indicates a slight bias to longer outputs, however, the difference is still small.