How Does India Cook *Biryani*? – Supplementary Material

C V Rishi[†]
IIIT Hyderabad
India
se23ucse044@mahindrauniversity.edu.in

Farzana S[†]
IIIT Hyderabad
India
farzana.s@research.iiit.ac.in

Shubham Goel[†]
IIIT Hyderabad
India
shubham.goel@students.iiit.ac.in

Aditya Arun
IIIT Hyderabad
India
adityaarun1@gmail.com

C V Jawahar
IIIT Hyderabad
India
jawahar@iiit.ac.in

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Scene understanding**; **Activity recognition and understanding**; *Video segmentation.*

Keywords

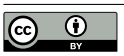
Video Understanding, Vision Language Models

ACM Reference Format:
C V Rishi, Farzana S, Shubham Goel, Aditya Arun, and C V Jawahar. 2025. How Does India Cook *Biryani*? – Supplementary Material. In *Indian Conference on Computer Vision, Graphics, and Image Processing (ICVGIP 2025), December 17–20, 2025, Mandi, India.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3774521.3774596

S1 - Biryani Categories

(1) **Ambur Biryani:** Originating from Ambur in Tamil Nadu, this version is made with flavorful seeraga samba (Jeeraga Samba) rice, lending a distinct aroma and texture. It is said to have royal roots in the Nawab of Arcot's kitchens and is typically served with a tangy eggplant curry.
(2) **Bombay Biryani:** A fusion of Persian, Mughlai, and Maharashtrian styles, Bombay biryani is a flavorful dum-cooked rice dish commonly featuring potatoes and sometimes dried plums, with a milder spice profile.
(3) **Dindigul Biryani:** Known particularly as Dindigul Thalapakatti Biryani, it uses seeraga samba rice and bold, tangy flavours—often featuring goat meat. It distinguishes itself through its slow-cooking technique and intense taste profile.
(4) **Donne Biryani:** A fragrant South Indian biryani, especially from Bangalore's "Military Hotel" style, this biryani uses seeraga samba rice and a freshly ground masala paste, served traditionally on a disposable leaf-paper "donne."
(5) **Hyderabadi Biryani:** Hailing from Hyderabad's Nizam kitchens, this iconic dum-cooked biryani comes in two variants: kachchi (raw marinated meat layered with rice) and pakki (cooked meat). It features basmati rice, meat, spices, saffron, and fried onions.
(6) **Kashmiri Biryani:** Typically a vegetarian style from Kashmiri Pandit tradition, it's made without onion or garlic and often includes vegetables, yoghurt, nuts, and fragrant basmati rice—a milder, saffron-infused version. Alternatively, mutton-based Kashmiri biryani includes dry fruits and kewra for a delicate flavour.
(7) **Kolkata Biryani:** Invented in the 1850s–60s by Nawab Wajid Ali Shah in exile, this biryani incorporated potatoes, eggs, lightly spiced meat, and fragrant rice, adapted from Awadhi-like Mughlai cooking—a lighter, more economical version.
(8) **Lucknow Awadhi Biryani:** From Lucknow's royal kitchens, this biryani is known for its subtle, fragrant flavours, often enhanced with kewra/rose water and saffron. It uses cooked meat layered with al dente rice and steamed "in dum" for refinement.
(9) **Malabar Biryani:** A signature of Kerala's Malabar coast (Kozhikode, Kannur, etc.), this subtly spiced biryani uses short-grain Kaima (Jeerakasala) rice, aromatic ghee, and whole spices like cardamom and cinnamon. It's mildly sweet, layered with fried onions, cashews, raisins, and cooked on dum for a fragrant, balanced flavour.
(10) **Mughlai Biryani:** Rooted in Mughal royal cuisine, this biryani is lavish and indulgent—made with basmati rice, meat (or vegetarian), cream, nuts, dried fruits, saffron, and aromatic spices, layered and dum-cooked for rich, creamy indulgence.
(11) **Sindhi Biryani:** A spicy, tangy, and sweet Pakistani biryani from Sindh, it includes potatoes, tomatoesyoghurtrt, dried plums (aloo bukhara), and a medley of spices. It's layered and dum-cooked, known for its bold, vibrant flavours.
(12) **Thalassery Biryani:** A celebrated local variant from Thalassery in North Kerala, this pakki-style biryani separately cooks Kaima rice and meat, then layers them for slow dum cooking. It's known for its dry, aromatic profile—no oil-heavy richness—and distinctive Kerala spices and ghee-infused rice.

[†]Equal Contribution



ICVGIP 2025, Mandi, India

ACM ISBN 979-8-4007-1930-1/2025/12
https://doi.org/10.1145/3774521.3774596

S2 - Prompts Used

Video Segmentation

InternVL-14B Prompt for Segment Analysis

You are analysing a cooking video.
Please extract information into three clearly labelled bullet-point lists, based strictly on what is visually present in the video frames.
Respond only with the following three sections in this exact order:
Ingredients: - List all visible ingredients being used (e.g., chopped onions, turmeric powder, rice).
Utensils: - List all visible cooking tools, vessels, or utensils (e.g., knife, pressure cooker, ladle).
Actions: - Describe each distinct cooking action as a verb-noun phrase (e.g., chopping onions, frying spices, stirring curry).
Important rules: - Do NOT include any summary, explanation, or extra commentary. - Only include items that are visible or implied in the visuals. - Avoid repeating the same item unless used in a different context. - Use consistent and specific terms.

Clustering Decision Prompt

You are analysing cooking actions for a biryani recipe classifier. Below is a set of len(actions) similar cooking actions that have been grouped:
actions_str
Question: Should these actions be split into multiple distinct action classes, or are they similar enough to remain as one group?
Consider: - Are there distinct cooking techniques or steps represented? - Would separating them improve classification accuracy for biryani cooking? - Are some actions fundamentally different despite semantic similarity?
Respond with a JSON object containing only: "should_split": true/false,

Gemini Prompt for Action Verification

You are an expert in analysing cooking videos. Your task is to determine if a specific action is happening in the provided video frames.
The action to verify is: **'[ACTION]'**
If any *part* of the action is clearly or partially visible—e.g., if the action is "adding turmeric and milk" but only turmeric is visible—answer "yes".
Only answer "no" if none of the described actions is visible. Do not explain. Respond with a single word: "yes" or "no".

Action Differencing Prompt

I am analysing two sets of photos (`{total_frames}` total) of someone performing the same biryani cooking action:

"`{action}`".
Video A: Photos `{clip1_range}`
Video B: Photos `{clip2_start}`–`{clip2_end}`

The specific difference to check is: "`{query_string}`".
This means I want to determine if Video A shows more of this characteristic compared to Video B.

`{importance_context}`

Question: Based on these frames, which video shows more of this difference?

(a) Video A
(b) Video B
(c) They look similar, or it's not clear
(d) The videos seem to be irrelevant to the query

Be careful: look at the entire set of frames for each video. If you are not confident or if the difference is very minor, choose (c).

Important Guidelines:

- Choose (a) if Video A clearly shows more of the difference than Video B
- Choose (b) if Video B clearly shows more of the difference than Video A
- Choose (c) if you cannot confidently distinguish between them or they appear similar
- Choose (d) if the videos do not relate to the query at all / the action shown is completely different to the cooking action

Return JSON:

```
{
    "answer": "a|b|c|d",
    "confidence": 1-5,
    "difference_visible": true/false,
    "explanation": "Detailed explanation
                    of what you observed"
}
```

QA Generation

The following prompts, templates, and illustrative examples present the full details of the input specifications used in our multi-stage question–answer (QA) generation pipeline. While the main paper outlines the methodology at a conceptual level, this section provides the exact instructions given to language models, along with representative intermediate outputs, to ensure reproducibility and transparency.

To produce segment-level natural language descriptions from 10-second video chunks, InternVL3-14B was guided with instructions emphasising explicit mention of ingredients, utensils, cooking actions, and other visually salient details, while avoiding speculative or unverifiable information.

Video Captioning Prompt

Generate a detailed and accurate description of a cooking video segment.
Use the following guidelines to craft a clear and complete narrative:

(1) Describe key visual elements such as ingredients, utensils, appliances, and the appearance of food at different stages of preparation.
(2) Focus on the sequence of actions performed by the cook, including preparation steps (e.g., chopping, mixing, frying), cooking techniques, and transformations in the food (e.g., colour changes, texture changes, boiling).
(3) Highlight interactions between the cook and the ingredients, as well as gestures or tools used.
(4) Emphasise the order of events, transitions between cooking stages, and any significant visual or temporal cues that indicate progress in the recipe.
(5) Ensure the description is thorough yet clear, capturing the essential visual and procedural aspects of the segment to help the viewer understand what is being cooked and how.

Figure 1 presents an example of a segment-level visual description generated by this captioning stage. The output demonstrates the desired level of detail and specificity, forming the foundation for subsequent summarisation and QA generation.

```
=== Chunk 1 ===
Start frame: 0
End frame: 300
Description:
The video showcases a dish involving rice and various ingredients being cooked and served onto
a plate. Initially, the video displays a plate of rice seasoned with saffron strands and
assorted vegetables and spices. As the video progresses, ingredients like fried pieces of
dough or patties, pieces of boiled and seasoned vegetables, and slices of papadum are added to
the rice. These elements are mixed into the rice, ensuring even distribution of flavors and
textures. The video concludes with a fully plated serving of the mixed rice dish, highlighting
the rich, golden color of the rice, the variety of vegetables, and the contrasting textures of
the fried patties and papadum.
================================================================================
=== Chunk 2 ===
Start frame: 300
End frame: 600
Description:
The video begins with an initial focus on a variety of ingredients laid out on a kitchen
counter, which include sliced chicken, potatoes, garlic, cilantro, chopped tomatoes, and white
rice among other ingredients. The camera then pans slightly to the left, revealing a container
with a blue lid, likely containing cooking oil or broth. A large orange bowl filled with white
rice is visible, suggesting the preparation of a rice dish. The ingredients are arranged
neatly in small bowls and plates, indicating the mise en place stage before the cooking
process. The emphasis in the frames is on showcasing the diverse array of ingredients rather
than any active cooking steps. The environment appears to be a home kitchen with a tiled
backsplash, indicating a setting prepared for a cooking session.
================================================================================
=== Chunk 3 ===
Start frame: 600
End frame: 900
Description:
The video showcases a variety of ingredients meticulously arranged on a wooden countertop,
signaling preparation for a substantial meal. In the arrangement, there is an orange bowl
filled with a thick white batter-like substance, and an adjacent large plastic jar with a blue
lid appearing to hold preserved foodstuffs. There are chunks of yellow potatoes and raw
chicken pieces neatly placed in separate plates. Various small bowls hold an assortment of
chopped vegetables, including green herbs, green bell peppers, tomatoes, and what seem to be
sliced mushrooms. Additionally, there are several spices in tiny bowls, likely including salt
and possibly other seasonings. Each item is clearly displayed, ready to be used in the cooking
process. The setting suggests a kitchen with tile walls in the background, indicating a clean
and organized cooking environment.
```

Figure 1: Video Description Example

For the next stage, Gemini-2.0-Flash was prompted to merge all chunk-level descriptions from a video into a coherent, temporally ordered summary. The instructions prioritised preserving event sequence, incorporating visually rich details, and eliminating redundancy, resulting in unified narratives suitable for downstream question generation.

Video Summarisation Prompt

We split a cooking video into segments and extracted detailed descriptions for each segment. The descriptions for all segments are listed below, in the order they appear in the video. For example, 'CHUNK: 1' corresponds to the first video segment.

Generate a detailed, step-by-step, and visually rich description of the entire cooking video as a single coherent paragraph, based on all the provided captions. Make sure not to lose any important information.
"""
<segment descriptions>
"""
Use the following instructions to create a clear, complete, and engaging cooking narrative:

(1) Focus on describing key visual details such as the appearance and colours of ingredients, textures, cooking methods, utensils used, hand movements, and how ingredients are combined or transformed during the process.
(2) Preserve the sequence of cooking actions — describe the preparation steps in the order they happen, ensuring the flow matches the progression shown in the captions.
(3) Highlight important details like quantities shown, specific types of ingredients (e.g., green chilli, rice, ginger garlic paste, potatoes), notable textures (e.g., moist, oily, tender), and garnishing or plating details.
(4) Use your reasoning to combine and organise information from all captions into one clear, thorough description. Remove unnecessary repetition and ignore any conflicting or irrelevant details.
(5) Do not mention that the information comes from captions. Present it as a natural, direct description of the video.
(6) Keep it visually descriptive yet easy to understand, almost like explaining the video to someone who can't watch it.
(7) Finally, use your common sense to conclude what dish is being prepared and summarise how the video showcases its preparation. If the video ends with plating or serving, describe that presentation too.

Figure 2 shows an example of a synthesised cooking-video summary produced from multiple segment descriptions. This illustrates how fragmented local observations are transformed into a continuous, recipe-level account.

The pipeline then included an information extraction step in which LLaMA-3-8B-Instruct identified three fixed categories — ingredients, utensils, and cooking actions — from a single segment description.

The cooking video begins with a close-up of a shiny metal pot where rice, thin orange cheese shreds, pieces of meat (likely chicken), and green chili peppers (both whole and sliced) are being stirred together with a big spoon, ensuring the ingredients are evenly mixed. The video then transitions to a close-up of a steaming pot filled with rice, chopped vegetables, and meat chunks, which is then served onto a metal plate alongside a fried egg. The preparation of the meat component of the dish is then shown, starting with 1 kg of raw red meat in a metal bowl. The meat is then transferred into a pressure cooker containing a seasoned liquid base, followed by the addition of two glasses of water. One teaspoon of salt is sprinkled over the meat, followed by a teaspoon of cumin seeds and red spice. Next, dried herbs (7/8 cloves), 1 tsp of black peppercorns, 4/5 green cardamom pods, and 2 teaspoons of dried fennel seeds are added to the meat. Parsley and garlic leaves, along with bay leaves, are also added. The pot is then covered with a lid and cooked. The pressure cooker lid, with visible steam condensation, indicates the cooking process is underway, and the contents are cooked until the pressure cooker emits 7/8 whistles on medium flame, revealing a stew of meat and possibly vegetables in a thick, dark broth. A ladle is then used to stir a pot filled with bones, meat, and other ingredients submerged in a rich, brown simmering liquid. Separately, oil is poured into another pot, followed by chopped yellow onions, which are then lightly fried with additional oil until they turn a light golden color. In a mixing bowl, a yellowish liquid is mixed with fresh green chili peppers and ginger garlic paste. In another pot, a noodle dish is stir-fried, incorporating noodles, green beans, sesame seeds, and a thick sauce, with water added to thin the sauce. Green peppers, garlic, and tomatoes are sautéed in a pot, followed by the addition of pieces of meat, which are stirred to ensure thorough mixing with the ingredients and liquid. Chunks of meat and sliced garlic are heated in a pot with oil, followed by a liquid, and left to boil, creating a mixture of green bell peppers and brown meat. A simmering pot of meat, green chili peppers, and amber-hued liquid is stirred with a utensil, and a portion of the meat is lifted to showcase its texture. A granular white substance is added to a large pot of stewed meat, green vegetables, and a yellowish broth, and stirred in. A hot soup featuring a meaty broth with chunks of meat and green chili peppers is stirred with a ladle.

Figure 2: Video Summary Example

Easy QA Generation Prompt

Video segment description:

```
"""
<segment description>
"""
```

Answer the following clearly:

(1) What are the ingredients shown in this segment?
(2) What are the utensils shown in this segment?
(3) What are the cooking actions performed in this segment?

Medium-difficulty QA relied on a curated set of question templates covering ingredient usage, step ordering, cooking durations, presentation details, and utensil usage, ensuring questions were grounded in observable visual evidence. These templates were combined with video summaries and transcripts, enabling Gemini-2.0-Flash to generate richer question–answer pairs that integrated multiple sources while avoiding irrelevant or speculative details.

Medium QA Templates

(1) **What are the primary ingredients used in this recipe?**
e.g., chicken, rice, yoghurt, spices, onions, tomatoes
(2) **In what order are the ingredients added during cooking?**
e.g., oil → spices → onions → meat → tomatoes → yogurt
(3) **Which spices or seasonings are used in this dish?**
e.g., cumin seeds, coriander powder, garam masala, turmeric, salt
(4) **What kind of meat is used in the recipe?**
e.g., goat, chicken, fish, lamb, beef, none
(5) **What is the first step shown in the video?**
e.g., rinsing and soaking the rice, marinating the meat
(6) **What is the last step before serving?**
e.g., garnishing with fresh coriander and fried onions
(7) **How is the meat prepared before cooking?**
e.g., marinated with yoghurt, turmeric, and chilli powder, layered with meat
(8) **What type of pan or vessel is used to cook this dish?**
e.g., a wide heavy-bottomed metal pot, clay pot, pressure cooker
(9) **How long is the rice cooked for?**
e.g., approximately 15 minutes until tender
(10) **Approximately how long does it take to prepare this entire dish?**
e.g., around 45 minutes total
(11) **What does the final dish look like?**
e.g., orange-red rice with chicken pieces and green garnish
(12) **What is used to garnish the dish before serving?**
e.g., chopped coriander leaves, fried onions, lemon slices
(13) **Does the dish appear to be spicy?**
e.g., yes, it looks spicy due to the visible rechillili oil
(14) **When is the rice mixed with the meat or gravy?**
e.g., after the meat is cooked for 15 minutes
(15) **Is the dish served with any accompaniments?**
e.g., onion raita, boiled eggs, salad

Below is the full prompt provided to Gemini-2.0-Flash for medium-level QA generation. The instructions integrate video summaries with audio transcripts, combine template-guided and model-generated questions, and require answers grounded in the complete cooking process.

Medium QA Generation Prompt

You are an expert in analysing cooking videos, with extensive knowledge of culinary techniques, ingredients, and food presentation across various regional cuisines in India.
You are provided with a detailed textual description of the cooking video and the full transcript of the spoken narration. This data includes step-by-step cooking processes, mentions of ingredients, utensils, cooking durations, and visual cues — but you do not have access to the actual video.

Task:

- Identify and describe the key cooking processes, ingredients, and presentation details discussed in the textual description

and summary. (The key cooking process refers to the main focus of the video that is highlighted in the provided text.)
- Generate relevant Question-Answer (QA) pairs by carefully analysing the textual description and summary of the cooking video.
- In addition to using the provided template questions, feel free to create additional QA pairs that are contextually appropriate based on the content.
Below is a set of template questions for forming QA pairs: (Adapt these or create new ones depending on the content.)

```
"""
<templates>
"""
```

Instructions:

- DO NOT mention the video summary or transcript directly when answering the questions. Avoid phrases like: "based on the description," "according to the text," "as mentioned," or references to captions that imply the answer was derived from the provided text. Instead, present the information as if it is directly inferred from watching the video.
- Do not explain or justify how the answer was obtained.
- You may choose to omit details that seem irrelevant to the cooking process or final dish.
- Keep all answers concise, and highlight important keywords using bold formatting.
- If a particular question does not apply to the video, simply do not generate a QA pair for it.
- Focus on content directly relevant to the cooking process, ingredients, or presentation. Ignore unrelated background commentary.

Output Format:

```
{
  "Summary": "",
  "QA_pairs": [
    {"Q": "", "A": ""},
    {"Q": "", "A": ""},
    {"Q": "", "A": ""},
    {"Q": "", "A": ""}
  ]
}
```

Video description:

```
"""
<video description>
"""
```

Transcript:

```
"""
<transcript>
"""
```

The next stage involved creating multimodal summaries by combining detailed visual descriptions with transcribed spoken instructions. These summaries captured both appearance and process details, incorporating cooking tips, quantities, and sequencing from the narration.

Multimodal Summarisation Prompt

We have split a cooking video into visual segments and extracted detailed descriptions from the video frames for each segment. Separately, we also generated a full transcript of the audio narration spoken in the video.

Your task is to produce a comprehensive, visually and verbally rich summary of the entire cooking video by carefully combining information from both the visual descriptions and the audio transcript.

Video description from visual frames:

```
"""
<video description>
"""
```

Transcript of the audio narration:

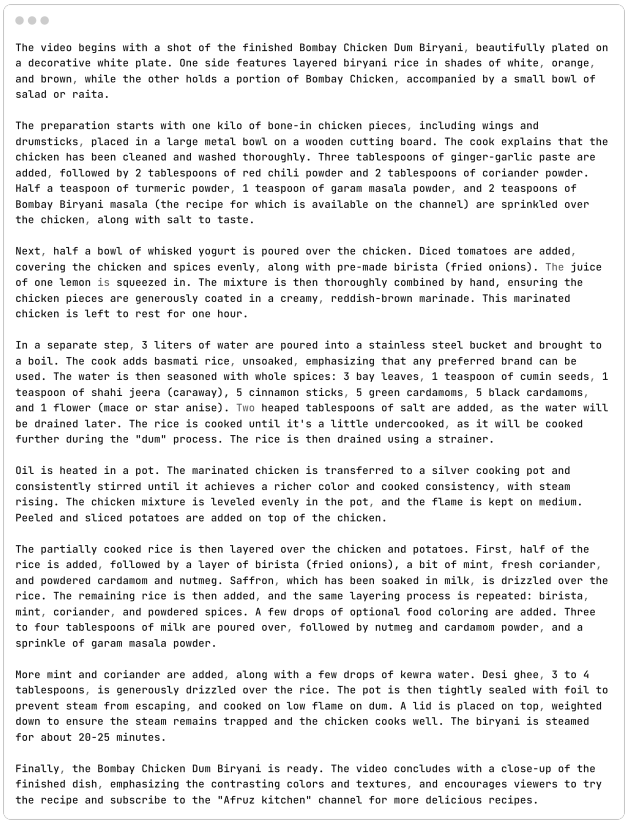
```
"""
<transcript>
"""
```

Use the following instructions to create a clear, complete, and engaging cooking video summary:

(1) Use the video summaries from frames to describe key visual details such as the appearance and colours of ingredients, textures, cooking methods, utensils, hand movements, how ingredients are layered or transformed, and plating or serving scenes.
(2) Use the transcript of the audio narration to incorporate spoken explanations, cooking tips, quantities, and verbal emphasis on techniques or ingredient choices.
(3) Ensure the cooking steps are described in the correct sequence, matching the flow shown across the video segments and the spoken instructions.
(4) Highlight important specifics like ingredient types (e.g., green chillies, basmati rice, ginger garlic paste, bone-in chicken), notable textures (e.g., golden fried onions, oily masala, tender meat), quantities or approximate amounts mentioned, and final garnishing or plating details.
(5) Merge and organise all this information into one clear, thorough, and engaging description, removing unnecessary repetition and ignoring conflicting or irrelevant details.
(6) Do not mention captions, transcripts, or segments explicitly. Present it as if you are naturally describing what is happening in the video.
(7) Keep the narrative vivid and easy to understand, as if explaining the video to someone who cannot watch it.
(8) Conclude by summarising what dish is being prepared and how the video showcases its preparation, including the final presentation if shown.

Figure 3 provides an example of such a multimodal summary, illustrating how complementary visual and auditory information is integrated into a single, highly detailed representation of the cooking process.

Finally, reasoning-intensive QA generation was carried out by comparing and contrasting multiple multimodal summaries. A dedicated set of high-level question templates supported cross-video



The video begins with a shot of the finished Bombay Chicken Dum Biryani, beautifully plated on a decorative white plate. One side features layered biryani rice in shades of white, orange, and brown, while the other holds a portion of Bombay Chicken, accompanied by a small bowl of salad or raita.

The preparation starts with one kilo of bone-in chicken pieces, including wings and drumsticks, placed in a large metal bowl on a wooden cutting board. The cook explains that the chicken has been cleaned and washed thoroughly. Three tablespoons of ginger-garlic paste are added, followed by 2 tablespoons of red chili powder and 2 tablespoons of coriander powder. Half a teaspoon of turmeric powder, 1 teaspoon of garam masala powder, and 2 teaspoons of Bombay Biryani masala (the recipe for which is available on the channel) are sprinkled over the chicken, along with salt to taste.

Next, half a bowl of whisked yogurt is poured over the chicken. Diced tomatoes are added, covering the chicken and spices evenly, along with pre-made birista (fried onions). The juice of one lemon is squeezed in. The mixture is then thoroughly combined by hand, ensuring the chicken pieces are generously coated in a creamy, reddish-brown marinade. This marinated chicken is left to rest for one hour.

In a separate step, 3 liters of water are poured into a stainless steel bucket and brought to a boil. The cook adds basmati rice, unsoaked, emphasizing that any preferred brand can be used. The water is then seasoned with whole spices: 3 bay leaves, 1 teaspoon of cumin seeds, 1 teaspoon of shahi jeera (caraway), 5 cinnamon sticks, 5 green cardamoms, 5 black cardamoms, and 1 flower (mace or star anise). Two heaped tablespoons of salt are added, as the water will be drained later. The rice is cooked until it's a little undercooked, as it will be cooked further during the "dum" process. The rice is then drained using a strainer.

Oil is heated in a pot. The marinated chicken is transferred to a silver cooking pot and consistently stirred until it achieves a richer color and cooked consistency, with steam rising. The chicken mixture is leveled evenly in the pot, and the flame is kept on medium. Peeled and sliced potatoes are added on top of the chicken.

The partially cooked rice is then layered over the chicken and potatoes. First, half of the rice is added, followed by a layer of birista (fried onions), a bit of mint, fresh coriander, and powdered cardamom and nutmeg. Saffron, which has been soaked in milk, is drizzled over the rice. The remaining rice is then added, and the same layering process is repeated: birista, mint, coriander, and powdered spices. A few drops of optional food coloring are added. Three to four tablespoons of milk are poured over, followed by nutmeg and cardamom powder, and a sprinkle of garam masala powder.

More mint and coriander are added, along with a few drops of kewra water. Desi ghee, 3 to 4 tablespoons, is generously drizzled over the rice. The pot is then tightly sealed with foil to prevent steam from escaping, and cooked on low flame on dum. A lid is placed on top, weighted down to ensure the steam remains trapped and the chicken cooks well. The biryani is steamed for about 20-25 minutes.

Finally, the Bombay Chicken Dum Biryani is ready. The video concludes with a close-up of the finished dish, emphasizing the contrasting colors and textures, and encourages viewers to try the recipe and subscribe to the "Afruz kitchen" channel for more delicious recipes.

Figure 3: Multimodal Summary Example

analysis, addressing similarities and differences in ingredients, techniques, spice usage, preparation order, and presentation styles. This stage required synthesis across multiple examples to produce challenging, reasoning-oriented question–answer pairs.

Hard QA Templates

(1) **Which ingredient is common across all the recipes shown?**
e.g., onions are used in all three dishes
(2) **Which dish uses the highest variety of spices?**
e.g., the Hyderabad biryani uses 7 different spices, more than the others
(3) **Which recipe takes the longest time to prepare?**
e.g., the Lucknow biryani takes approximately 1 hour
(4) **Which of the recipes do not include yoghurt as an ingredient?**
e.g., only the Ambur biryani skips yoghurt
(5) **In which video is rice boiled separately before adding to the meat, unlike in the others?**
e.g., the Lucknow recipe
(6) **Which recipe appears thspiciestcy?**
e.g., the Andhra biryani looks deep red from heavy chilli usage
(7) **In which video does the cook add the meat later in the cooking process compared to the others?**
e.g., the Kerala biryani adds meat after vegetables
(8) **Which videos are the most different from each other?**
e.g., the Kerala and Hyderabad biryanis differ greatly in cooking method and garnish
(9) **Which videos are the most similar to each other?**
e.g., the Ambur and Tamil Nadu biryanis are nearly identical

Below is the final prompt used with Gemini-2.5-Flash to generate reasoning-intensive QA pairs requiring the integration of information from multiple multimodal video summaries. It instructs the model to identify and synthesise cross-video patterns and distinctions that cannot be inferred from a single source.

Hard QA Generation Prompt

You are an expert in analysing cooking videos, with extensive knowledge of culinary techniques, ingredients, and food presentation across various regional cuisines in India.
You are provided with textual summaries of multiple cooking videos. These summaries include step-by-step actions, mentions of ingredients, utensils, and visual cues — but you do not have access to the actual videos themselves.

Task:

- Carefully compare, contrast, and synthesise the details across these multiple videos to identify key differences, similarities, and unique aspects. This includes analysing cooking processes, ingredients, preparation times, spice usage, visual appearance, and sequencing of steps.
- Generate high-level, challenging Question-Answer (QA) pairs that require reasoning across these multiple videos, not just describing a single video.
- Use the provided set of question templates to guide your QA generation. You may also create additional multi-video QA pairs if they are insightful.

Below is a set of template questions for forming QA pairs: (Adapt these or create new ones depending on the content.)

```
"""
<templates>
"""
```

Instructions:

- Do not mention the video summaries or textual descriptions directly when answering the questions. Avoid phrases like: "based on the description," "according to the text," "as mentioned," or references to captions that imply the answer was derived from the provided summaries. Instead, present the information as if it is directly inferred from watching the videos.
- Do not explain or justify how the answer was obtained.
- Keep all answers concise, and highlight important keywords using bold formatting.
- If a particular question does not apply to this set of videos, simply do not generate a QA pair for it.
- Focus on content directly relevant to the cooking processes, ingredients, or comparative aspects. Ignore unrelated background commentary.

Output Format:

```
{
  "Summary": "",
  "QA_pairs": [
    {"Q": "", "A": ""},
    {"Q": "", "A": ""},
    {"Q": "", "A": ""},
    {"Q": "", "A": ""}
  ]
}
```

Video summaries:

```
"""
<video summaries>
"""
```

S3 - Question Answer Examples

This section presents representative question–answer (QA) pairs from the easy, medium, and hard difficulty tiers of the dataset. These examples illustrate how the prompts, templates, and generation procedures described in Section S2 are applied in practice, highlighting the distinct characteristics and reasoning demands of each difficulty level.

The easy tier focuses on localised, segment-level visual observations. Questions are designed to be direct and unambiguous, answerable from a short video segment without requiring broader temporal or cross-modal reasoning.

Figures 4–6 showcase three easy-tier examples, each containing concise, factual questions about ingredients, utensils, or cooking actions visible within a specific segment.

The medium tier integrates information from entire video summaries and transcripts. These questions require temporal sequencing, recognition of ingredient roles, and interpretation of the overall cooking process.


```
[
  {
    "video": "ambur_biryani_video7",
    "chunk": 45,
    "start_frame": 13200,
    "end_frame": 13500,
    "question": "What are the utensils shown in this segment?",
    "answer": "Pink non-slip grip silicone pot holder"
  },
  {
    "video": "ambur_biryani_video7",
    "chunk": 59,
    "start_frame": 17400,
    "end_frame": 17700,
    "question": "What are the ingredients shown in this segment?",
    "answer": "Golden brown rice, Red and green vegetables, Chunks of meat, Boiled eggs"
  }
]
```


Figure 4: Easy Example 1


```
[
  {
    "video": "hyderabadi_biryani_video8",
    "chunk": 11,
    "start_frame": 6000,
    "end_frame": 6600,
    "question": "What are the ingredients shown in this segment?",
    "answer": "Raw meat, Green chilies, Bay leaves, Cloves, Cinnamon stick, Ground spices,
Ginger paste, Butter, Fresh leaves, Lemon juice, Fried shallots"
  },
  {
    "video": "hyderabadi_biryani_video8",
    "chunk": 20,
    "start_frame": 11400,
    "end_frame": 12000,
    "question": "What are the utensils shown in this segment?",
    "answer": "A copper measuring cup, A stainless steel spoon"
  }
]
```


Figure 5: Easy Example 2


```
[
  {
    "video": "lucknow_awadhi_biryani_video4",
    "chunk": 56,
    "start_frame": 16500,
    "end_frame": 16800,
    "question": "What are the ingredients shown in this segment?",
    "answer": "Biryani rice, Chunks of meat, Sliced carrots, Whole green chilies, Green herbs
(for garnish)"
  },
  {
    "video": "lucknow_awadhi_biryani_video4",
    "chunk": 6,
    "start_frame": 1500,
    "end_frame": 1800,
    "question": "What are the cooking actions performed in this segment?",
    "answer": "Stirring, scooping, lifting, and checking for doneness/tenderness"
  }
]
```


Figure 6: Easy Example 3

Figures 7–9 illustrate medium-tier examples, where answering requires synthesising information across multiple steps of preparation while remaining grounded in observable content.

The hard tier requires multi-video comparative and contrastive reasoning. These questions cannot be answered from a single video alone; they demand integration of information across multiple cooking demonstrations to identify similarities, differences, and unique patterns.

Figures 10–13 present four examples from this tier, demonstrating reasoning over ingredient variations, cooking methods, spice usage, preparation order, and presentation styles across different recipes.

```
 {
    "video": "ambur_biryani_video9",
    "question": "What are the primary ingredients used in this recipe?",
    "answer": "The primary ingredients are chicken, seeraga samba rice, onions, tomatoes,
ginger-garlic paste, red chilies, yogurt, and various spices."
  },
  {
    "video": "ambur_biryani_video9",
    "question": "In what order are the ingredients added during cooking?",
    "answer": "The ingredients are added in the following order: bay leaf, clove, cardamom →
onions → tomatoes, mint, coriander → ginger-garlic paste → chili paste → curd → lemon
→ chicken → rice → water."
  },
  {
    "video": "ambur_biryani_video9",
    "question": "Which spices or seasonings are used in this dish?",
    "answer": "The spices and seasonings used are bay leaf, clove, cardamom, red chilies, and
salt."
  },
  {
    "video": "ambur_biryani_video9",
    "question": "What kind of meat is used in the recipe?",
    "answer": "Chicken is used in the recipe."
  },
  {
    "video": "ambur_biryani_video9",
    "question": "How is the meat prepared before cooking?",
    "answer": "The chicken is mixed with a masala consisting of chili paste, curd, and lemon."
  }
```


Figure 7: Medium Example 1


```
  {
    "video": "dindigul_biryani_video3",
    "question": "What type of pan or vessel is used to cook this dish?",
    "answer": "A kadai (pan) is used to sauté the masala and roast the mutton, and a pressure
cooker is used to cook the mutton initially."
  },
  {
    "video": "dindigul_biryani_video3",
    "question": "How long is the rice cooked for?",
    "answer": "The rice is initially cooked on medium flame for about 10 minutes, then goes on
'dum' for 15 minutes."
  },
  {
    "video": "dindigul_biryani_video3",
    "question": "Approximately how long does it take to prepare this entire dish?",
    "answer": "The biryani can be made quickly compared to other types."
  },
  {
    "video": "dindigul_biryani_video3",
    "question": "What is used to garnish the dish before serving?",
    "answer": "The final dish appears to be garnished with yogurt or cream, a fresh green
sprig, potato slices, and a boiled egg."
  },
  {
    "video": "dindigul_biryani_video3",
    "question": "What is the ratio of rice to water used in the recipe?",
    "answer": "The rice-to-water ratio is 1:2."
  }
```


Figure 8: Medium Example 2


```
  {
    "video": "mughlai_biryani_video6",
    "question": "What is the first step shown in the video?",
    "answer": "The first step shown is measuring and washing long-grain basmati rice and
soaking it in water."
  },
  {
    "video": "mughlai_biryani_video6",
    "question": "How is the meat prepared before cooking?",
    "answer": "The meat is marinated with spices, ginger-garlic paste, saffron milk, and
yogurt."
  },
  {
    "video": "mughlai_biryani_video6",
    "question": "Approximately how long does it take to cook the biryani on low flame?",
    "answer": "The biryani is cooked on low flame for 45 minutes."
  },
  {
    "video": "mughlai_biryani_video6",
    "question": "What is used to garnish the dish before serving?",
    "answer": "The dish is garnished with fried chicken, a fried egg, sliced red onions,
sliced yellow squash, sliced cashews, green chili peppers, and fresh green coriander leaves."
  },
  {
    "video": "mughlai_biryani_video6",
    "question": "What other ingredients are mixed with the rice?",
    "answer": "The rice is mixed with dal, cinnamon, mace, star anise, coriander leaves, mint
leaves, bay leaves and nutmeg powder."
  }
```


Figure 9: Medium Example 3


```
  {
    "videos": [
      "kashmiri_biryani_video5",
      "bombay_biryani_video10"
    ],
    "question": "Which ingredient is common across both recipes?",
    "answer": "Both recipes commonly use salt, oil, basmati rice, onions, yogurt, coriander,
turmeric, garam masala, cinnamon, cloves, and cardamom."
  },
  {
    "videos": [
      "kashmiri_biryani_video5",
      "bombay_biryani_video10"
    ],
    "question": "Which dish uses the highest variety of spices?",
    "answer": "The Bombay Biryani includes a broader array of spices, such as whole cumin
seeds, black peppercorns, nutmeg, mace powder, and dried plums, in addition to common biryani
spices."
  },
  {
    "videos": [
      "kashmiri_biryani_video5",
      "bombay_biryani_video10"
    ],
    "question": "Which recipe takes the longest time to prepare?",
    "answer": "The Bombay Biryani requires a two-hour marination period for the mutton and at
least an hour of slow cooking for the meat, making it the more time-intensive recipe."
  }
```


Figure 10: Hard Example 1

S4 - Evaluation Metrics

BLEU

The Bilingual Evaluation Understudy (BLEU) metric is an algorithm used to assess the quality of text generated by machine translation from one natural language to another. Its core principle is that the closer a machine's translation is to that of a skilled human translator, the higher its quality. Developed at IBM in 2001, BLEU was among the first metrics to demonstrate a strong correlation with human quality judgments and remains a widely used, low-cost automatic evaluation method.

BLEU computes scores for individual translated segments—typically sentences—by comparing them against one or more high-quality reference translations. These segment-level scores are then averaged across the entire corpus to estimate overall translation quality. The metric does not account for intelligibility or grammatical accuracy.

The BLEU score ranges from 0 to 1, with higher values indicating greater similarity to the reference translations. A score of 1 is rare even among human translations, as it requires an exact match with a reference. Consequently, a perfect score is not necessary to indicate high quality.

ROUGE-L

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a widely adopted framework for evaluating the quality of automatically generated summaries—and occasionally translations—by measuring their similarity to one or more human reference texts. The resulting scores range from 0 to 1, with higher values indicating greater alignment with the references.

Among the ROUGE variants, ROUGE-L distinguishes itself by leveraging the Longest Common Subsequence (LCS) between the candidate and reference texts, thereby capturing sentence-level structural similarity rather than merely local n-gram matches. It calculates recall as the ratio of LCS length to the total length of the

```
  {
    "videos": [
      "hyderabadi_biryani_video5",
      "mughlai_biryani_video7",
      "thalassery_biryani_video3"
    ],
    "question": "Which two biryani recipes are the most similar in their overall cooking
approach and choice of rice?",
    "answer": "The Hyderabadi Biryani and Mughlai Biryani are the most similar, both primarily
using basmati rice and involving separate boiling and layering of rice, followed by a dum
cooking method."
  },
  {
    "videos": [
      "hyderabadi_biryani_video5",
      "mughlai_biryani_video7",
      "thalassery_biryani_video3"
    ],
    "question": "Which biryani recipe stands out as most distinct from the others in terms of
its spice preparation, rice type, and use of traditional cookware?",
    "answer": "The Thalassery Biryani is the most distinct, characterized by its unique dry-
roasting and grinding of whole spices, use of Jeera Samba rice, and the prominent use of clay
pots and a stone mortar and pestle for preparation."
  }
```


Figure 11: Hard Example 2


```
{
    "videos": [
      "mughlai_biryani_video3",
      "ambur_biryani_video7",
      "dindigul_biryani_video2",
      "mughlai_biryani_video6"
    ],
    "question": "Which videos are the most different from each other?",
    "answer": "The Mughlai Chicken Dum Biryani and the Dindigul Thalappakatti Chicken Biryani
show the most significant differences in their approach to rice cooking, spice preparation,
and core ingredients."
  },
  {
    "videos": [
      "mughlai_biryani_video3",
      "ambur_biryani_video7",
      "dindigul_biryani_video2",
      "mughlai_biryani_video6"
    ],
    "question": "Which videos are the most similar to each other?",
    "answer": "The two Mughlai biryani preparations are the most similar, both utilizing pre-
cooked basmati rice, layering techniques, saffron, and a sealed dum method."
  }
```


Figure 12: Hard Example 3


```
{
    "videos": [
      "kashmiri_biryani_video4",
      "kashmiri_biryani_video7",
      "mughlai_biryani_video1",
      "bombay_biryani_video7",
      "kolkata_biryani_video2"
    ],
    "question": "Which recipe requires the longest preparation time due to an extended
marination period?",
    "answer": "The Mughlai Biryani takes the longest to prepare, requiring an extensive
marination period of at least two hours or even overnight."
  },
  {
    "videos": [
      "kashmiri_biryani_video4",
      "kashmiri_biryani_video7",
      "mughlai_biryani_video1",
      "bombay_biryani_video7",
      "kolkata_biryani_video2"
    ],
    "question": "Which of the recipes do not include yogurt as a direct ingredient?",
    "answer": "Neither of the Kashmiri biryani recipes includes yogurt."
  }
```


Figure 13: Hard Example 4

reference, precision as the ratio of LCS length to the total length of the candidate, and combines these measures via an F_1 score.

ROUGE-L's ability to reward the preservation of word order and coherence makes it particularly useful for assessing the structural fidelity of condensed text. For instance, even when individual words match, a summary with a disrupted sequence will receive a lower ROUGE-L score compared to one that maintains the original flow, highlighting its sensitivity to sentence structure.

BERTScore

BERTScore is an advanced evaluation metric introduced in 2019 for assessing the quality of machine-generated text by leveraging contextual embeddings derived from pre-trained models like BERT. Unlike traditional evaluation methods such as BLEU or ROUGE, which rely on surface-level word or n-gram matching, BERTScore evaluates semantic similarity through token-level cosine similarity in the embedding space.

The mechanism operates by embedding each token of both the candidate and reference texts using a BERT-based model. It then computes the cosine similarity between all token pairs, using a greedy matching strategy: each candidate token aligns with the most semantically similar reference token for precision, and vice versa for recall. These scores are then harmonised into an F1 measure; optional enhancements such as inverse document frequency (IDF) weighting or baseline rescaling can be applied.

Empirical validation has shown that BERTScore correlates more strongly with human evaluations across various text generation tasks—such as machine translation, summarisation, and image captioning—than traditional metrics. It is particularly effective at capturing semantic equivalence in cases involving paraphrasing or lexical variation.

By focusing on contextual understanding rather than exact token overlap, BERTScore provides a more nuanced and human-aligned evaluation of generated language, making it especially valuable in modern NLP and generative model assessments.

S5 - Video Segmentation

Action clustering

Direct application of InternVL-14B across thousands of segments yields detailed action descriptions that often vary lexically despite being semantically identical. To address this redundancy, we employed an agglomerative clustering with average linkage on action phrase embeddings generated using the all-MiniLM-L6-v2 SentenceTransformer model. We used a cosine distance of 0.3 to merge clusters; will no pairs fall below this threshold, we then pick a representative phrase to be the action label.

This clustering process significantly reduces the action vocabulary while preserving semantic diversity.

The initial action detection stage produced a highly granular label space with 10,481 unique action classes. After applying the action clustering process, this number was reduced to 2,187 canonicalised action classes, representing a 79.1% reduction while greatly improving consistency in labelling.

Temporal Merging

To further enhance temporal coherence, we implemented a clip merging procedure to address fragmentation where identical actions span consecutive temporal segments. This temporal merging process significantly reduced fragmentation in the video segmentation. Across all videos, the number of timestamped clips decreased from 16,761 before merging to 14,479 after merging, representing

a 13.6% reduction in segment count while preserving full action coverage.

Table 1: Action clustering and temporal merging statistics showing significant consolidation in both label space and temporal segmentation

Process	Before	After	Reduction (%)
Action clustering	10,481 classes	2,187 classes	79.1
Temporal merging	16,761 clips	14,479 clips	13.6

Example Data Representation

To illustrate how our dataset is structured, we provide two representative JSON snippets. The first shows a **10-second temporal segment** annotated with ingredients, utensils, and actions. The second shows an **action-to-timestamp mapping**, where semantically similar action phrases are clustered, and each cluster contains all associated video clips.

10-second Segment Annotation


```
{
  "timestamp": "59-69",
  "title": "Hyderabadi Chicken Dum Biryani #biryani",
 "url": "https://www.youtube.com/watch?v=BIXMwLFCboA&t=59s",
  "ingredients": [
    "Mint Leaves",
    "Coriander Leaves",
    "Kesar Milk",
    "Kewra & Rose Water",
    "Ghee"
  ],
  "utensils": [
    "Large cooking pot or bowl",
    "Orange cup",
    "Metal cup"
  ],
  "actions": [
    "Adding mint leaves to rice",
    "Adding coriander leaves to rice",
    "Pouring kesar milk over rice",
    "Pouring kewra and rose water over rice",
    "Pouring ghee over rice"
  ]
}
```


Action-to-Timestamped Clips Mapping


```
"adding bay leaves to the grinder": {
  "phrases": [
    "adding bay leaves to the grinder",
    "placing bay leaf in the spice grinder"
  ],
  "clips": [
    {
    "url": "https://www.youtube.com/watch?v=hgI4wV_WoVs&t=80s",
      "timestamp": "80-90",
      "biryani": "dindigul_biryani",
      "video": "video10"
    },
    {
    "url": "https://www.youtube.com/watch?v=5Zra4nFepRg&t=139s",
      "timestamp": "139-149",
      "biryani": "dindigul_biryani",
      "video": "video1"
    }
  ]
}
```


These structured annotations enable fine-grained temporal localisation of cooking actions, association with relevant ingredients and utensils, and grouping of semantically similar actions across different videos. This organisation supports multimodal reasoning tasks such as step retrieval, ingredient localisation, and cross-video action comparison.

Verification Workflow

We compile candidate segments grouped by canonical action (e.g., "marinating chicken," "adding whole spices"), each stored with metadata for action label, video URL, local file path, timestamps (in seconds), *biryani* type, and video index. For each 10–30 s segment, we sample up to 20 evenly spaced RGB frames using OpenCV to ensure temporal coverage while controlling input size. These frames are paired with a structured natural language prompt asking Gemini to confirm whether the specified action occurs, where partial or incomplete visibility counts as valid evidence. We query Gemini 2.5 Flash Lite with low temperature for deterministic yes/no outputs, then parse responses as *Correct* for "Yes," *Incorrect* for "No," and *Error* for ambiguous or API failures.

Implementation Details

The complete video segmentation pipeline was executed on NVIDIA A40 GPUs with 48GB VRAM, requiring approximately 12 hours of computation time. InternVL-14B [**?**] processed 14,470 video segments across all *biryani* varieties, while the clustering phase operated on the resulting action embeddings using scikit-learn's agglomerative clustering implementation [**?**].

S7 - Video Comparison Results

Implementation Details

Our video comparison framework processed comparisons across 12 *biryani* varieties based on clustered action classes (Table 1). Since action classes contain multiple video instances, the number of pairwise comparisons grows as $\binom{n}{2}$ where n is the number of clips per action class. Popular action classes like "stirring" (348 instances) and "stirring/mixing rice" (210 instances) (Table 3) generated substantially more comparisons than smaller classes.

Table 2: Implementation details for video segmentation pipeline components showing computational requirements and processing scope

Component	Model	Processing Scope	Compute Requirements
Action detection	InternVL-14B	16,761 video segments	NVIDIA A40 (48GB)
Action clustering	SentenceTransformer	10,481 unique actions	CPU-based
Temporal merging	Rule-based	16,761 → 14,479 clips	CPU-based
Verification	Gemini 2.5 Flash Lite	14,479 merged segments	Google API

Table 3: Top action classes by instance count from clustering results

Action Class	Instances
stirring	348
stirring/mixing rice	210
pouring rice and liquid	169
placing/removing pressure cooker lid	142
scooping rice and ingredients	134
stirring pot contents	130
preparing onions	127
mixing ingredients in the pot	125
serving the *biryani*	112
assembling chicken and rice	107
stirring/adding chicken	106
stirring the mixture	102

The Proposer stage (Qwen2.5) ran once per action class to generate plausible variations. The Frame Localizer (CLIP with ViT-BigG-14) processed every clip instance within each action class. Both components operated on NVIDIA A40 GPUs with 48GB VRAM, requiring approximately 40 hours each. The Action Differencer used Gemini 2.5 Flash Lite in batch processing mode for final comparisons.

Regional Variation Analysis

Cross-regional comparisons reveal consistent patterns where certain cooking stages maintain similarity across *biryani* types while others exhibit substantial variation. For each pairwise regional comparison (*Hyderabadi* vs *Kolkata*, *Hyderabadi* vs *Lucknowi*, etc.), fundamental preparation chapters remain consistent while specific execution stages diverge based on cultural techniques.

Comparison Statistics

The framework detected differences in 33.2% of total comparisons. This percentage represents comparison-level detection: if any proposed difference within a comparison pair was identified, the entire comparison was counted as "difference detected." A comparison was marked as having differences even if only one of multiple proposed variations was found.

If measuring absolute difference detection rather than comparison-level detection, the rate would be approximately 19%, reflecting the granular nature of individual variation identification within each comparison.

For manual verification accuracy assessment, we use individual difference detection, counting each specific proposed difference separately. For manual verification, we want to know how well our model performed rather than how varied our data is.

Future Improvements

The framework's limitations suggest specific enhancement directions:

- **Enhanced Proposer knowledge**: Deeper understanding of Indian cooking techniques would enable generation of more comprehensive difference categories, particularly when processing large clip volumes per action class.
- **Fine-tuned visual encoding**: CLIP's general training may miss fine-grained cooking actions specific to Indian culinary contexts. Increasing retrieved frame counts or specialised model fine-tuning could improve detection granularity.

Despite current limitations, the framework successfully captures meaningful procedural differences across regional *biryani* varieties, providing systematic insights into traditional cooking method diversity.

Table 4: Implementation details for video comparison framework components

Component	Model	Processing Scope	Compute Requirements
Proposer	Qwen2.5	Once per action class	NVIDIA A40 (48GB), ~40 hours
Frame Localizer	CLIP ViT-BigG-14	Every clip instance	NVIDIA A40 (48GB), ~40 hours
Action Differencer	Gemini 2.5 Flash Lite	Pairwise comparisons	Batch processing mode through the Gemini API

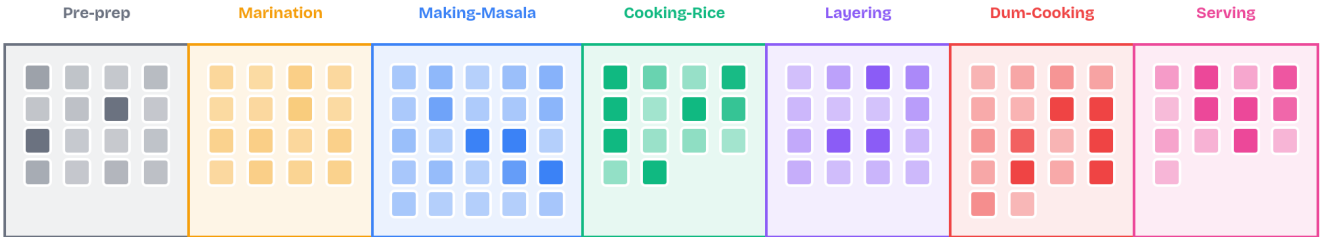


Figure 14: *Hyderabadi biryani* vs *Kolkata biryani* variation visualization. Node opacity indicates the degree of detected procedural differences across cooking stages.