A. YouCookQA dataset

In this section, we introduce the detail of how our question-answer (QA) pairs and alternative choices are collected, together with some examples in our dataset. Then, we show some more statistics of the dataset.

QA collection For the first step of collection, the workers at Amazon Mechanical Turk (AMT) are asked to propose one question of specified type about the video, and provide an answer according to the video content. By proposing the correct answers together with questions, we want to make sure that the questions are answerable by watching the video, rather than collecting arbitrary unsolvable questions. Each of the six categories of questions are collected separately to ensure that each video can have at least one QA pair for each category. Specifically for the "multi-hop" category, which is a more general one that may overlap with other types of questions, we design a check-box interface for workers to decide which specific type(s) the QA pairs belong to, as a result which, some questions are tagged by multiple categories.

QA demonstration In Tab. 1, we show a representative QA pair for each category. We bold the keywords to illustrate why a specific QA pair is categorized as that type. E.g., in "time" category, the question requests for comparing the speed, which is related to time; In "multi-hop" category, we have multiple steps of reasoning (two bold words). Also, the demo QA pair for "multi-hop" category is marked as "order" category because it has "last/before" keywords which are related to order information.

Alternative collection After the QA pairs being collected, we collect additional alternatives to compose multiple choice questions. Human workers are asked to annotate alternatives and clean the multiple choice QA following the three rules: 1) The alternatives need to be related to the question; 2) The alternatives need to have different meanings from the correct answer; 3) If the correct answer is not reasonable or too subjective for a machine to answer, edit the answer or delete the QA. Through this step, the QA pairs collected in the first step are double-checked. Alternatives for "how many" questions are generated randomly with integers ranging from 0 to 10 except for the correct answer, and those for "when" questions are random numbers within the range of $[0, A_{Correct}-50] \cup [A_{Correct}+50, 600]$, where $A_{Correct}$ is the number in the correct answer. If $A_{Correct}$ is less than 50, then we just keep the right interval of the range, or if $A_{Correct}$ is greater than 550, we keep the left one. 600 is set as a heuristic upper bound of the answers to "when" questions in our dataset, since 95.8% of the videos in our dataset is less than 600 seconds.

Extra statistics In Table. 2, we collect the total number of QA pairs, and statistics for four tags of different answer formats. Also, in 3, we show the statistics of QA pairs per video, number of categories per QA pair, and average an-

Table 1: Demonstration of QA pairs of six different categories. YTID stands for YouTube ID. The bold words are the keywords by which the questions are treated as specific categories. Correct answers of the multiple choice questions are given as marked with ✓.

Categories	YTID	Examples
Count	GLd3aX16zBg	Q) **How many** pieces is the sandwich cut into?
		A) 6. B) 1. C) 2.✓ D) 4. E) 3.
Order	XOwypmUT5cc	Q) Which ingredient is added **first**?
		A) Sesame seeds.✓ B) Tomatoes. C) Garlic. D) Marinara sauce. E) Alfredo sauce.
Taste	pOWe4zB-E-4	Q) Is the masala dosa **spicy**?
		A) 423. B) No. C) 273. D) 308. E) Yes.✓
Time	K6Uk5vNi1_Q	Q) Is salt added **faster** the first or second time?
		A) The first time B) The fifth time. C) Not added. D) The sixth time. E) The second time.✓
Multi-hop	xHr8X2Wpmno	Q) Is salt the **last** ingredient to be added to the bowl **before** lettuce is chopped?
		A) 45. B) Tahini. C) Yes. D) No.✓ E) 157.
Property	Gs3OGfQbPjc	Q) What **color** are the onion rings after being fried?
		A) Green. B) Brown.✓ C) Red. D) Gray. E) Black.

Table 2: Statistics for each QA type

	Total	Yes/No	Num	Single	Text
Total	15355	3241	6188	1501	4425
Count	3980	550	3065	189	174
Order	3260	653	348	495	1761
Taste	2000	1574	11	92	322
Time	2950	29	2885	3	31
Multi-hop	7200	868	2363	1349	2612
Property	3267	411	84	716	2052

Table 3: Statistics for each QA type

	QA/video	Tags/QA	Ans avg len
Count	1.990	1.625	1.187
Order	1.630	1.490	3.925
Taste	1.000	1.003	2.039
Time	1.475	1.403	1.074
Multi-hop	3.600	2.021	2.130
Property	1.634	2.016	3.047

swer length (by number of words in the answer) with respect to six different QA categories.

B. Human quiz setup

Apart from using deep learning models to complete VideoQA tasks, we also invite ten human annotators to perform human test. The participants are asked to imitate the procedure of training and testing of a neural network. A general picture is that each person is given a batch of 300 QA pairs selected from the training set for reference. Afterwards, they will take a quiz of 30 QA pairs from the test set. The QA pairs for both training and testing are randomly sampled according to the distribution of 6 question tags. The main difference between human beings and machines is that human beings can make use of their powerful common sense to complete the task. In detail, the game consists of three steps.

Common sense The first step is called "Common sense". The participants need to answer the multiple-choice questions without watching the videos. The training set is purely QA pairs without video, where participants can explore answer patterns. Also, random guess and common sense are used in this step.

Visual only In this step, muted videos are available to participants as visual information. This is because in our baseline models, only visual features are used. The participants will complete the same 30 multiple-choice questions again.

Visual&Audio (CC) Now, audio is turned on. Transcripts (CC) are allowed in case the audio is not clear enough. The main reason why we do not use audio in other models is that the audio tracks of all the cooking videos are mainly speeches, which are already converted to CC. Also, the collection of our dataset does not involve audio information, so our dataset does not contain any questions involving sounds made by utensils or ingredients. Again, under this setting, the participants will redo the quiz.

Finally, all three sets of answers for the three steps are collected and analyzed together with results from the deep learning models. In Tab. 4, we list the detailed accuracy scores of the human quiz.

C. Full results

In Tab. 5, 6, we show all the experiment results, part of which are omitted in the main body of the paper. Apart from multiple choice accuracy, K-Space is another evaluation metric for our dataset, where a MLP with K output neurons is tasked to predict the correct answer by using Eq. 2: $\mathrm{A}^* = \arg\max_{j=1,\ldots,K} g_j(v, q)$.

D. Visualization

In our attached video, we show several example results on different models, and make the comparison among the results. Detailed analysis is provided in the video.

Table 4: Human quiz accuracy scores.

Quiz steps	Count	Order	Taste	Time	Multi-hop	Property	Total
Common sense	0.535	0.432	0.654	0.485	0.511	0.588	0.528
Visual only	0.973	0.975	0.936	0.990	0.988	0.982	0.970
Visual&Audio (CC)	0.973	0.975	0.966	0.990	0.988	0.982	0.977

Table 5: Multiple choice accuracy scores.

Models	Modalities	Count	Order	Taste	Time	Multi-hop	Property	Total
Baselines	Bare QA	0.435	0.321	0.466	0.239	0.292	0.438	0.348
	Naive RNN	0.434	0.330	0.467	0.234	0.283	0.449	0.347
	MAC	0.438	0.331	0.462	0.229	0.294	0.437	0.348
SEQ	Visual	0.452	0.337	0.476	0.230	0.288	0.449	0.352
	Description	0.453	0.350	0.463	0.244	0.285	0.443	0.353
	Transcript	0.454	0.320	0.469	0.227	0.286	0.438	0.347
	V+CC	0.450	0.328	0.468	0.241	0.293	0.432	0.351
	V+D	0.450	0.328	0.468	0.241	0.293	0.432	0.351
SEQ-SA	Visual	0.455	0.337	0.465	0.238	0.298	0.447	0.355
	Description	0.449	0.334	0.473	0.231	0.286	0.419	0.347
	Transcript	0.450	0.331	0.472	0.241	0.299	0.453	0.357
	V+CC	0.448	0.337	0.471	0.241	0.306	0.461	0.361
	V+D	0.448	0.337	0.471	0.241	0.306	0.461	0.361
GCN	Visual	0.452	0.341	0.464	0.224	0.282	0.427	0.346
	Description	0.452	0.331	0.475	0.229	0.290	0.447	0.352
	Transcript	0.451	0.337	0.470	0.251	0.282	0.416	0.348
	V+CC	0.449	0.317	0.475	0.223	0.287	0.455	0.349
	V+D	0.449	0.317	0.475	0.223	0.287	0.455	0.349
GCN-SA	Visual	0.477	0.343	0.487	0.229	0.311	0.446	0.365
	Description	0.449	0.334	0.473	0.231	0.286	0.419	0.347
	Transcript	0.521	0.374	0.469	0.245	0.310	0.449	0.375
	V+CC	0.516	0.383	0.481	0.280	0.309	0.461	0.383
	V+D	0.516	0.383	0.481	0.280	0.309	0.461	0.383
RGCN	Visual	0.522	0.371	0.478	0.277	0.329	0.490	0.392
	Description	0.514	0.370	0.482	0.274	0.314	0.483	0.385
	Transcript	0.496	0.348	0.473	0.240	0.298	0.466	0.361
	V+CC	0.518	0.350	0.480	0.269	0.325	0.493	0.390
	V+D	0.536	0.399	0.497	0.286	0.334	0.513	0.413
RGCN-SA	Visual	0.513	0.388	0.483	0.289	0.336	0.498	0.403
	Description	0.516	0.379	0.483	0.286	0.313	0.493	0.389
	Transcript	0.484	0.343	0.469	0.243	0.319	0.485	0.366
	V+CC	0.516	0.373	0.471	0.271	0.326	0.497	0.393
	V+D	0.523	0.398	0.494	0.286	0.337	0.516	0.416

Table 6: K-Space accuracy scores.

Models	Modalities	Count	Order	Taste	Time	Multi-hop	Property	Total
Baselines	Bare QA	0.183	0.132	0.153	0.111	0.103	0.183	0.141
	Naive RNN	0.189	0.135	0.155	0.109	0.102	0.188	0.145
	MAC	0.172	0.124	0.149	0.103	0.097	0.180	0.138
SEQ	Visual	0.190	0.152	0.173	0.133	0.124	0.193	0.160
	Description	0.201	0.151	0.165	0.124	0.117	0.199	0.158
	Transcript	0.188	0.147	0.172	0.126	0.119	0.200	0.159
	V+CC	0.193	0.143	0.173	0.124	0.123	0.193	0.151
	V+D	0.194	0.149	0.168	0.137	0.124	0.199	0.160
SEQ-SA	Visual	0.206	0.153	0.175	0.121	0.127	0.204	0.164
	Description	0.187	0.147	0.178	0.129	0.117	0.177	0.156
	Transcript	0.183	0.150	0.166	0.122	0.111	0.183	0.152
	V+CC	0.191	0.160	0.173	0.132	0.127	0.197	0.167
	V+D	0.204	0.163	0.184	0.143	0.133	0.198	0.173
GCN	Visual	0.184	0.144	0.169	0.126	0.120	0.188	0.150
	Description	0.196	0.152	0.165	0.124	0.127	0.201	0.157
	Transcript	0.176	0.131	0.156	0.115	0.113	0.187	0.143
	V+CC	0.187	0.141	0.167	0.128	0.110	0.177	0.150
	V+D	0.183	0.136	0.154	0.115	0.108	0.186	0.148
GCN-SA	Visual	0.196	0.155	0.176	0.134	0.124	0.199	0.164
	Description	0.199	0.140	0.165	0.126	0.112	0.193	0.153
	Transcript	0.180	0.136	0.172	0.119	0.119	0.194	0.150
	V+CC	0.207	0.170	0.181	0.140	0.134	0.206	0.177
	V+D	0.223	0.173	0.190	0.153	0.143	0.224	0.183
RGCN	Visual	0.216	0.168	0.193	0.145	0.134	0.223	0.179
	Description	0.204	0.152	0.178	0.134	0.136	0.201	0.163
	Transcript	0.193	0.146	0.160	0.127	0.127	0.193	0.152
	V+CC	0.205	0.159	0.186	0.143	0.136	0.203	0.173
	V+D	0.213	0.180	0.204	0.160	0.151	0.227	0.194
RGCN-SA	Visual	0.226	0.164	0.199	0.153	0.140	0.216	0.182
	Description	0.199	0.156	0.174	0.135	0.122	0.195	0.162
	Transcript	0.183	0.134	0.153	0.110	0.107	0.197	0.144
	V+CC	0.213	0.172	0.190	0.151	0.139	0.225	0.180
	V+D	0.243	0.196	0.214	0.172	0.170	0.233	0.203