

933 **Supplementary Material**

934 **6 Prompts Used**

935 Here are some of the prompts we developed for generating EgoSchema.

936 **6.1 Set A**

937 **6.1.1 Question prompt**

938 Input:

939 I want you to act as a teacher in the class called "Long-term video understanding".

940 I will provide video action narrations and their timestamps and you will
941 generate three highly difficult and diverse questions for your students about
942 the high-level details in the video. You want to test students' following
943 abilities:

944 Ability 1: Students' ability to summarize and compare long parts of the video

945 Ability 2: Students' ability to compress information from the video rather than just
946 listing the actions that happened in the video.

947 Ability 3: Students' ability to identify the most important parts of the video.

948

949 Your questions should not mention any particular timestamps or narrations. Remember
950 to make sure the correct answers to your questions do not list information from
951 the narrations but compress them in a concise conclusion.
952

953

954 Examples of good and difficult questions:

955 "What is the main action of the video?"

956 "Why did C do action ...?"

957

958 AVOID the following types of questions:

959 "When ...?"

960 "How many ...?"

961 "How much ...?"

962

963 When announcing the question please label each question as "Question 1,2,3: [full
964 question]"

965

966

967 Timestamps and narrations:

968 0 - C stares at the lamp

969 1 - C looks around the apartment

970 3 - C talks to a man

971 6 - C walks around the apartment

972 10 - C talks to a man

973 11 - C looks around the apartment

974 14 - a man plays the guitar

975 20 - a man walks around the apartment

976 21 - a man plays the guitar

977 21 - C walks around the apartment

978 22 - C stares at the window

979 24 - C looks around the apartment

980 27 - C talks to a man

981 29 - C picks chips from the table

982 32 - C looks around the apartment

983 36 - C stares at a man

984 36 - a man plays the guitar

985 37 - C walks around the apartment

986 41 - C sits on the sofa

987 43 - C stares at a man

988 44 - a man plays the guitar

989 49 - a man climbs up the sofa

990 50 - a man plays the guitar

991 62 - a man climbs down the sofa

992 63 - a man walks around the apartment
993 64 - a man plays the guitar
994 93 - C stares at the table
995 96 - a man plays a guitar
996 111 - C stares at the table
997 112 - a man walks around the apartment
998 114 - C stares at a man
999 115 - a man lifts the guitar
1000 118 - a man walks around the apartment
1001 120 - a man places guitar on the sofa
1002 122 - a man walks around the apartment
1003 125 - C stands up
1004 126 - C walks around the apartment
1005 131 - a man picks a coffee maker jug
1006 132 - a man pours coffee in a cup
1007 134 - C stares at a man
1008 140 - C stares at the window
1009 142 - C stares at a man
1010 144 - a man drinks coffee
1011 145 - C looks around the apartment
1012 150 - C walks around the apartment
1013 152 - C talks to a man
1014 156 - C stares at a bench
1015 158 - C looks around the apartment
1016 164 - C stares at a man
1017 164 - a man talks to C
1018 167 - C looks around the apartment
1019 171 - C follows a man
1020 174 - a man points at the window
1021 175 - C looks around the bedroom
1022 =====
1023 Output:
1024 Question 1: What can you deduce about the relationship between C and the man, based
1025 on their interactions and behaviors throughout the video?
1026
1027 Question 2: Identify the primary activity in the apartment and discuss its
1028 importance/significance. How does it influence the actions and atmosphere of
1029 the video?
1030
1031 Question 3: Observe the dynamics of the video in terms of changes in characters'
1032 actions, interactions, and spatial movement. How do these shifts contribute to
1033 the overall narrative?

1034 6.1.2 Answer prompt

1035 Input:
1036 I want you to act as a teacher in the class called "Long-term video understanding."
1037 I will provide video action narrations and their timestamps and three highly
1038 difficult and diverse questions for your students about the high-level details
1039 in the video. I want you to test students' following abilities:
1040
1041 Ability 1: Students' ability to summarize and compare long parts of the video
1042 Ability 2: Students' ability to compress information from the video rather than just
1043 listing the actions that happened in the video.
1044 Ability 3: Students' ability to identify the most important parts of the video.
1045
1046 I want you to create a difficult multiple-choice exam that tests above student
1047 abilities based on the three questions I just provided. Each question should
1048 have five similar open-ended but short answers, but only one should be correct.
1049 Make it very difficult for students to find the correct answer among all the
1050 wrong answers. All answers should be closely related to what happens in the
1051 video. Make wrong answers significantly longer than correct answers. Ensure all
1052 of the correct answers compress information from narrations them into a

1053 concise conclusion. Your answers should not mention any particular timestamps
1054 or narrations.
1055
1056 Do not use letters for the answer choices
1057 Print each correct answer exactly as "Correct answer: [full answer]"
1058 Please print each wrong answer on a new line and print each wrong answer as "Wrong
1059 answer 1,2,3,4: [full answer]"
1060
1061
1062 Timestamps and narrations:
1063 0 - C stares at the lamp
1064 1 - C looks around the apartment
1065 3 - C talks to a man
1066 6 - C walks around the apartment
1067 10 - C talks to a man
1068 11 - C looks around the apartment
1069 14 - a man plays the guitar
1070 20 - a man walks around the apartment
1071 21 - a man plays the guitar
1072 21 - C walks around the apartment
1073 22 - C stares at the window
1074 24 - C looks around the apartment
1075 27 - C talks to a man
1076 29 - C picks chips from the table
1077 32 - C looks around the apartment
1078 36 - C stares at a man
1079 36 - a man plays the guitar
1080 37 - C walks around the apartment
1081 41 - C sits on the sofa
1082 43 - C stares at a man
1083 44 - a man plays the guitar
1084 49 - a man climbs up the sofa
1085 50 - a man plays the guitar
1086 62 - a man climbs down the sofa
1087 63 - a man walks around the apartment
1088 64 - a man plays the guitar
1089 93 - C stares at the table
1090 96 - a man plays a guitar
1091 111 - C stares at the table
1092 112 - a man walks around the apartment
1093 114 - C stares at a man
1094 115 - a man lifts the guitar
1095 118 - a man walks around the apartment
1096 120 - a man places guitar on the sofa
1097 122 - a man walks around the apartment
1098 125 - C stands up
1099 126 - C walks around the apartment
1100 131 - a man picks a coffee maker jug
1101 132 - a man pours coffee in a cup
1102 134 - C stares at a man
1103 140 - C stares at the window
1104 142 - C stares at a man
1105 144 - a man drinks coffee
1106 145 - C looks around the apartment
1107 150 - C walks around the apartment
1108 152 - C talks to a man
1109 156 - C stares at a bench
1110 158 - C looks around the apartment
1111 164 - C stares at a man
1112 164 - a man talks to C
1113 167 - C looks around the apartment
1114 171 - C follows a man
1115 174 - a man points at the window
1116 175 - C looks around the bedroom
1117

1118 Questions:
1119
1120 Question: what can you deduce about the relationship between c and the man, based on
1121 their interactions and behaviors throughout the video?
1122
1123 Question: identify the primary activity in the apartment and discuss its importance/
1124 significance. how does it influence the actions and atmosphere of the video?
1125
1126 Question: Observe the dynamics of the video in terms of changes in characters'
1127 actions, interactions, and spatial movement. how do these shifts contribute to
1128 the overall narrative?
1129 =====
1130 Output:
1131 Question 1: What can you deduce about the relationship between C and the man, based
1132 on their interactions and behaviors throughout the video?
1133
1134 Correct answer: They have a casual, friendly relationship.
1135 Wrong answer 1: C and the man are complete strangers and never interacted before.
1136 Wrong answer 2: C is deeply in love with the man, and the man reciprocates those
1137 feelings.
1138 Wrong answer 3: They are business associates discussing a financial transaction.
1139 Wrong answer 4: C is an intruder in the man's apartment, and the man is trying to
1140 get C to leave.
1141
1142 Question 2: Identify the primary activity in the apartment and discuss its
1143 importance/significance. How does it influence the actions and atmosphere of
1144 the video?
1145
1146 Correct answer: The primary activity is the man playing the guitar, which sets a
1147 relaxed environment.
1148 Wrong answer 1: The main activity is C staring at different objects, indicating a
1149 deep curiosity.
1150 Wrong answer 2: The primary activity is a heated argument between C and the man that
1151 escalates over time.
1152 Wrong answer 3: The main activity is the man anxiously pacing around the apartment,
1153 creating a tense atmosphere.
1154 Wrong answer 4: The primary activity is C and the man planning to commit a crime
1155 together.
1156
1157 Question 3: Observe the dynamics of the video in terms of changes in characters'
1158 actions, interactions, and spatial movement. How do these shifts contribute to
1159 the overall narrative?
1160 Correct Answer: Actions and interactions are casual and relaxed, reflecting a
1161 comfortable environment.
1162 Wrong Answer A : The video displays a profound sense of conflict and tension arising
1163 between the characters.
1164 Wrong Answer B : The man is showing C the issues that need fixing in the apartment
1165 in a professional manner.
1166 Wrong Answer C : Both the characters display an increasingly urgent need to solve an
1167 issue in the apartment.
1168 Wrong Answer D : C and the man admire and interact with several objects in the
1169 apartment that look beautiful.

1170 6.2 Set B

1171 6.2.1 Question and answer prompt

1172 Input:
1173 I want you to act as a teacher in the class called "Long-term video understanding".
1174 I will provide video action narrations and their timestamps and you will
1175 generate three highly difficult and diverse questions for your students about
1176 the high-level details in the video. You want to test students' following
1177 abilities:
1178

1179 Ability 1: Students' ability to summarize and compare long parts of the video
1180 Ability 2: Students' ability to compress information from the video rather than just
1181 listing the actions that happened in the video.
1182 Ability 3: Students' ability to identify the most important parts of the video.
1183
1184 Your questions should not mention any particular timestamps or narrations. Remember
1185 to make sure the correct answers to your questions do not list information from
1186 the narrations but compress them in a concise conclusion.
1187
1188 Examples of good and difficult questions:
1189 "What is the main action of the video?"
1190 "Why did C do action ...?"
1191
1192 AVOID the following types of questions:
1193 "When ...?"
1194 "How many ...?"
1195 "How much ...?"
1196
1197 When announcing the question please label each question as "Question 1,2,3: [full
1198 question]"
1199
1200
1201 Timestamps and narrations:
1202 3 - C holds the cloth in his right hand.
1203 5 - the woman picks a carton from the grocery bag on the floor with her right hand.
1204 6 - the woman drops the carton in a cabinet with her left hand.
1205 7 - the woman dips both hands into the grocery bag.
1206 9 - the woman drops a green carton on the floor with her right hand.
1207 12 - C drops the green carton in the cabinet with his right hand.
1208 13 - the woman holds a pack bag in her right hand.
1209 16 - C opens a kitchen cabinet with his left hand.
1210 18 - C removes a cereal pack from the kitchen cabinet with his left hand.
1211 19 - C puts the green carton into the kitchen cabinet with his right hand.
1212 21 - C closes the kitchen cabinet with his left hand.
1213 24 - the woman removes a plastic from the grocery bag with her right hand.
1214 25 - the woman drops the plastic on the floor with her right hand.
1215 33 - C closes a wardrobe with his left hand.
1216 38 - the woman puts a pack into the cabinet with her right hand.
1217 43 - a dog lies down on a bed.
1218 54 - C picks a cloth from the floor with his right hand.
1219 58 - C adjusts the cloth with both hands.
1220 66 - C hangs the cloth on the wall with both hands.
1221 74 - the woman holds a grocery bag in her right hand.
1222 82 - the woman touches her hair with her right hand.
1223 92 - the woman talks with C.
1224 99 - C holds two piece of cloths in both hands.
1225 100 - the woman picks piece of clothes from a bag with both hands.
1226 100 - C adjusts a camera on his head with his right hand.
1227 103 - C drops the two piece of cloths on a couch with his left hand.
1228 109 - C opens a door with his right hand.
1229 110 - C walks into a toilet.
1230 114 - C holds a red towel in his right hand.
1231 116 - the woman talks with C.
1232 118 - the woman scratches her back with her right hand.
1233 118 - the dog walks into the toilet.
1234 120 - C hangs the red towel on the wall with his right hand.
1235 121 - C holds a towel in his left hand.
1236 129 - the woman holds the towel in her left hand.
1237 129 - the dog walks out of the toilet.
1238 132 - C holds the door with his left hand.
1239 133 - the woman touches her face with her right hand.
1240 148 - C adjusts the camera on his head with his left hand.
1241 149 - C touches his face with his right hand.
1242 150 - the woman holds a bin with both hands.
1243 160 - C opens the door with his right hand.

1244 161 - C walks into the toilet.
1245 163 - the woman walks into the toilet.
1246 166 - the woman drops the bin on the floor with both hands.
1247 168 - the woman adjusts the bin on the floor with both hands.
1248 =====
1249 Output:
1250 Question 1: Can you provide a concise summary of the primary activities taking place
1251 in the video, highlighting similarities between the actions of C and the woman?
1252
1253
1254 Question 2: Based on the video, what task seems to be the central focus of both C
1255 and the woman's actions, and how do they contribute to the accomplishment of
1256 this task?
1257
1258 Question 3: Identify the three most important segments of the video and explain the
1259 significance of each in the overall context of the actions taking place.
1260 =====
1261 Input:
1262 Students are done with the questions. It is time to reveal the correct answers.
1263 Correctly provide short one-sentence answers to the questions you created, and
1264 additionally generate 4 very similar but wrong answers. Make sure that it will
1265 be hard to find the correct answer when combined with the wrong ones. Also,
1266 ensure that all the answers are concise and very close to an actual answer in
1267 length. Mainly make sure your correct answers do not list information from the
1268 narrations but compress them in a concise conclusion.
1269
1270 Do not use letters for the answer choices
1271 Print each correct answer exactly as "Correct answer: [full answer]"
1272 Please print each wrong answer on a new line and print each wrong answer as "Wrong
1273 answer 1,2,3,4: [full answer]"
1274 =====
1275 Output:
1276 Question 1:
1277 Correct answer: C and the woman both interact with various objects to organize the
1278 space, including placing grocery items and hanging cloths.
1279 Wrong answer 1: [Output disregarded]
1280 Wrong answer 2: [Output disregarded]
1281 Wrong answer 3: [Output disregarded]
1282 Wrong answer 4: [Output disregarded]
1283
1284 Question 2:
1285 Correct answer: The central focus is organizing and tidying up the space, with C and
1286 the woman putting items in their respective places.
1287 Wrong answer 1: [Output disregarded]
1288 Wrong answer 2: [Output disregarded]
1289 Wrong answer 3: [Output disregarded]
1290 Wrong answer 4: [Output disregarded]
1291
1292 Question 3:
1293 Correct answer: Key segments include organizing groceries, adjusting household items
1294 (e.g., cloths, bins), and interacting with a dog.
1295 Wrong answer 1: [Output disregarded]
1296 Wrong answer 2: [Output disregarded]
1297 Wrong answer 3: [Output disregarded]
1298 Wrong answer 4: [Output disregarded]

1299 **6.2.2 Wrong answer prompt**

1300 Input:
1301 I want you to act as a teacher in the class called "Long-term video understanding."
1302 I will provide video action narrations and their timestamps and three highly
1303 difficult and diverse questions for your students about the high-level details
1304 in the video. I will also show the correct answers to the questions. I want you
1305 to test students' following abilities:

1306
1307 Ability 1: Students' ability to summarize and compare long parts of the video
1308 Ability 2: Students' ability to compress information from the video rather than just
1309 listing the actions that happened in the video.
1310 Ability 3: Students' ability to identify the most important parts of the video.
1311
1312 I want you to create a difficult multiple-choice exam that tests above student
1313 abilities based on the three questions and their correct answers I just
1314 provided. Each question should have five similar open-ended but short answers,
1315 but only one should be correct. Make it very difficult for students to find the
1316 correct answer among all the wrong answers. All answers should be closely
1317 related to what happens in the video. Make wrong answers significantly longer
1318 than correct answers. Ensure all of the correct answers compress information
1319 from narrations them into a concise conclusion. Your answers should not mention
1320 any particular timestamps or narrations.
1321
1322 Do not use letters for the answer choices
1323 Please print each wrong answer on a new line and print each wrong answer as "Wrong
1324 answer 1,2,3,4: [full answer]"
1325
1326
1327 Timestamps and narrations:
1328 3 - C holds the cloth in his right hand.
1329 5 - the woman picks a carton from the grocery bag on the floor with her right hand.
1330 6 - the woman drops the carton in a cabinet with her left hand.
1331 7 - the woman dips both hands into the grocery bag.
1332 9 - the woman drops a green carton on the floor with her right hand.
1333 12 - C drops the green carton in the cabinet with his right hand.
1334 13 - the woman holds a pack bag in her right hand.
1335 16 - C opens a kitchen cabinet with his left hand.
1336 18 - C removes a cereal pack from the kitchen cabinet with his left hand.
1337 19 - C puts the green carton into the kitchen cabinet with his right hand.
1338 21 - C closes the kitchen cabinet with his left hand.
1339 24 - the woman removes a plastic from the grocery bag with her right hand.
1340 25 - the woman drops the plastic on the floor with her right hand.
1341 33 - C closes a wardrobe with his left hand.
1342 38 - the woman puts a pack into the cabinet with her right hand.
1343 43 - a dog lies down on a bed.
1344 54 - C picks a cloth from the floor with his right hand.
1345 58 - C adjusts the cloth with both hands.
1346 66 - C hangs the cloth on the wall with both hands.
1347 74 - the woman holds a grocery bag in her right hand.
1348 82 - the woman touches her hair with her right hand.
1349 92 - the woman talks with C.
1350 99 - C holds two piece of cloths in both hands.
1351 100 - the woman picks piece of clothes from a bag with both hands.
1352 100 - C adjusts a camera on his head with his right hand.
1353 103 - C drops the two piece of cloths on a couch with his left hand.
1354 109 - C opens a door with his right hand.
1355 110 - C walks into a toilet.
1356 114 - C holds a red towel in his right hand.
1357 116 - the woman talks with C.
1358 118 - the woman scratches her back with her right hand.
1359 118 - the dog walks into the toilet.
1360 120 - C hangs the red towel on the wall with his right hand.
1361 121 - C holds a towel in his left hand.
1362 129 - the woman holds the towel in her left hand.
1363 129 - the dog walks out of the toilet.
1364 132 - C holds the door with his left hand.
1365 133 - the woman touches her face with her right hand.
1366 148 - C adjusts the camera on his head with his left hand.
1367 149 - C touches his face with his right hand.
1368 150 - the woman holds a bin with both hands.
1369 160 - C opens the door with his right hand.
1370 161 - C walks into the toilet.

1371 163 - the woman walks into the toilet.
1372 166 - the woman drops the bin on the floor with both hands.
1373 168 - the woman adjusts the bin on the floor with both hands.
1374
1375 Questions:
1376
1377 Question: Can you provide a concise summary of the primary activities taking place
1378 in the video, highlighting similarities between the actions of c and the woman?
1379 Correct answer: c and the woman both interact with various objects to organize the
1380 space, including placing grocery items and hanging cloths.
1381
1382 Question: Based on the video, what task seems to be the central focus of both c and
1383 the woman's actions, and how do they contribute to the accomplishment of this
1384 task?
1385 Correct answer: the central focus is organizing and tidying up the space, with c and
1386 the woman putting items in their respective places.
1387
1388 Question: Identify the three most important segments of the video and explain the
1389 significance of each in the overall context of the actions taking place?
1390 Correct answer: key segments include organizing groceries, adjusting household items
1391 (e.g., cloths, bins), and interacting with a dog.
1392 =====
1393 Output:
1394 Question 1: Can you provide a concise summary of the primary activities taking place
1395 in the video, highlighting similarities between the actions of c and the woman?
1396
1397
1398 Correct answer: C and the woman both interact with various objects to organize the
1399 space, including placing grocery items and hanging cloths.
1400 Wrong answer 1: C and the woman engage in lengthy conversations about their day
1401 while occasionally placing groceries in cabinets and adjusting cloths on walls.
1402 Wrong answer 2: C and the woman focus on preparing an elaborate dinner, flipping
1403 through recipe books and chopping vegetables.
1404 Wrong answer 3: Both C and the woman predominantly pay attention to the dog's well-
1405 being, providing it with food and water and playing with it.
1406 Wrong answer 4: The video revolves around C and the woman working on a DIY project,
1407 with tools scattered around the room.
1408
1409 Question 2: Based on the video, what task seems to be the central focus of both c
1410 and the woman's actions, and how do they contribute to the accomplishment of
1411 this task?
1412
1413 Correct answer: The central focus is organizing and tidying up the space, with C and
1414 the woman putting items in their respective places.
1415 Wrong answer 1: C and the woman are preoccupied with an ongoing debate, their
1416 actions aimlessly directed towards unrelated tasks.
1417 Wrong answer 2: They are both primarily absorbed in their individual tasks,
1418 occasionally interacting with each other but with no clear common goal.
1419 Wrong answer 3: The central focus of their actions is providing care and attention
1420 to the dog, and their actions revolve around its needs.
1421 Wrong answer 4: The video primarily depicts a disagreement between C and the woman,
1422 with actions centered around this conflict.
1423
1424 Question 3: Identify the three most important segments of the video and explain the
1425 significance of each in the overall context of the actions taking place?
1426
1427 Correct answer: Key segments include organizing groceries, adjusting household items
1428 (e.g., cloths, bins), and interacting with a dog.
1429 Wrong answer 1: The main segments involve discussing a movie, checking on an
1430 overheating stove, and attempting to catch a rodent in the room.
1431 Wrong answer 2: The most important segments consist of C and the woman eating a meal
1432 , participating in a board game, and reading books.
1433 Wrong answer 3: Three primary segments include C and the woman debating household
1434 chores, searching for lost items, and tending to a crying baby.

1435 Wrong answer 4: Key segments include discussing an upcoming event, taking turns
1436 answering a phone call, and checking updates from an ongoing sports game.

1437 7 Our clip length and narration density choice

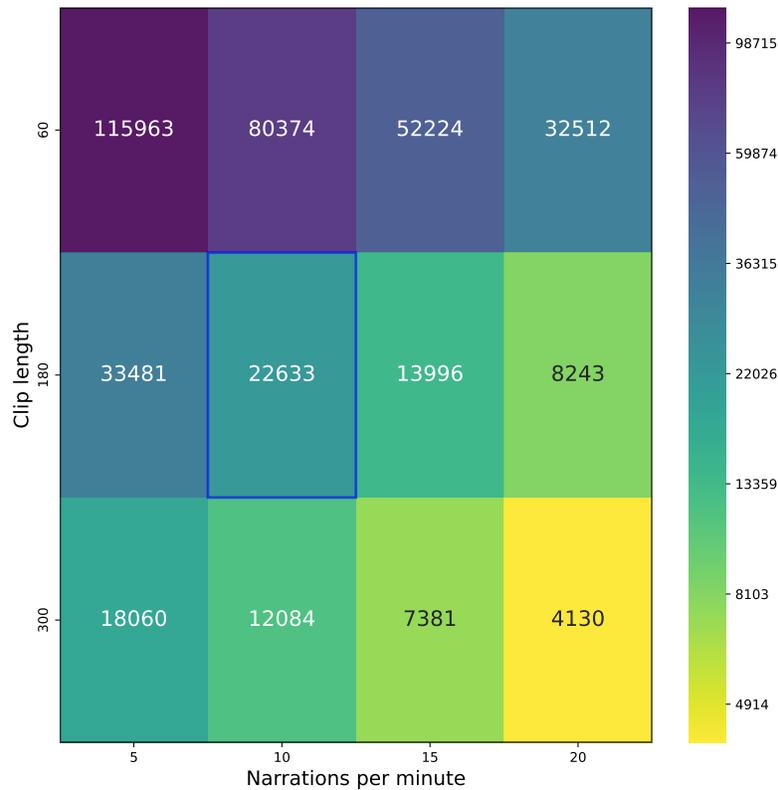


Figure 8: **Heatmap of number of viable clips over a range of clip length and narration density.** There are only a few viable options that offer some degree of balance between the number of clips and the number of narration in the clip. One potential selection is to utilize 3-minute clips with a density of 5 narrations per minute, although this choice bears the significant disadvantage of potentially including clips with an insufficient volume of narration data to generate high-quality results. Another possible choice is to use 1-minute clips with a density of 20 narrations per minute, yet this option carries the drawback of the clips being too brief for the dataset to be very long-term. Hence, we choose 3-minute clips with a narration density of 10 narrations per minute as it offers a satisfactory balance between the number of narrations and clip length for generating EgoSchema.

1438 8 Author Statement of Responsibility

1439 The authors confirm all responsibility in case of violation of rights. We will also ensure that
1440 EgoSchema is distributed under the Ego4D license.

1441 9 Hardware used

1442 All zero-shot benchmarking was done on the NVIDIA A100 Tensor Core GPU with 80 GB of
1443 memory. See Section 4 for the benchmarking results.

1444 10 Hosting, Licensing, and Long-Term Preservation Plan

1445 EgoSchema download instructions will be provided and explained on our [project page](#). The JSON
1446 file, which houses the textual component of the dataset, and the video clip download tool will be
1447 stored on the project’s GitHub page: <https://github.com/egoscema/EgoSchema>. Given GitHub’s
1448 widespread use and robust security measures, we are confident that this platform will ensure the
1449 stable accessibility and long-term preservation of our dataset. As for the licensing, EgoSchema will

1450 be publicly released under the Ego4D license, which allows direct public use of the video and text
1451 data for both research and commercial purposes.

1452 **11 Human curation**

1453 Our research utilized the services of Quantigo, a company specializing in data annotation. The
1454 process involved two distinct annotation procedures: data curation and human accuracy testing.

1455 **11.1 Curation**

1456 Generated data curation was performed by Quantigo employees. These curators were responsible for
1457 ensuring that the released EgoSchema dataset is the highest high quality possible. Here is the exact
1458 instructions that was provided to annotators:

1459 The annotation we need is to say that the Question-correct answer-wrong answer set (
1460 the whole set) is good if all these three conditions pass:

1461
1462 (Condition A) Question is Answerable: The question can be answered from the video
1463 and requires more than just a second of video to answer (so, if the answer is
1464 not present in the video or, if the answer can be formed with just a few frames
1465 (less than say, a second) then it fails this condition).
1466

1467 (Condition B) The Marked Correct Answer is correct: The "correct answer" is more
1468 the correct answer to the question
1469

1470 (Condition C) The Marked Wrong Answers are wrong: All 4 "wrong answers" are less
1471 correct than the "correct answer" (So for example, if a wrong answer is not
1472 completely false, but simply does not contain all the information that the "
1473 correct answer" does, then it is still a fine "wrong answer") IF even one of
1474 the marked answer is correct, the set should be labeled as bad.
1475

1476 (Condition D) The question is actually long-term: This is a very very important
1477 condition. We want the certificate for the question to be at least 30 seconds
1478 minimum. If the certificate is non-contiguous (ie. 5 seconds at one place, 20
1479 seconds at another, and 15 more seconds at a third place) the sum of lengths of
1480 all the sub-certificates together should be more than 30 seconds. Another
1481 example is, if a question can be answered simply from a few frames of the video,
1482 the certificate is small (and less than 30 seconds) and hence would fail this
1483 condition. Additional details on how to handle certificate edge cases are
1484 provided in the annotator training through examples.
1485

1486 (Condition E) Avoid Boring Questions: Questions that ask about the frequency of
1487 something ("How many times..") fail this condition.
1488

1489 If any of these five conditions fail we want the whole set (Question / Correct
1490 Answer / Wrong Answer) marked bad.

1491
1492 Optional:

1493 Since GOOD sets are so rare, in cases where it seems that a set is good but a small
1494 part of the above five conditions is not being met or, if one/two words were
1495 different this can be a good set, please label as MAYBE and we will fix it in
1496 the second round. We expect, Good/Bad to be about 97% of data and Maybe to be
1497 not more than 3%.
1498

1499 Extended notes:

1500 1. In our experience, the wrong answers are made such that they differ from the
1501 correct answer in small but crucially meaningful ways. There are many cases
1502 where a careless reading of the wrong answer might make it seem that it is
1503 correct but upon careful inspection, it will become clear that something about
1504 the wrong answer indeed makes it wrong. While this is common, there are indeed
1505 cases where the wrong answer is also just as correct as the correct answer. In
1506 such cases, the wrong answer fails condition C, and the set is bad.
1507

1508 2. Roughly speaking, we expect about 20-25% of the questions that we have provided
1509 to be found as good. However, this is not necessary and the %age can be smaller
1510 or larger depending on each three-minute clip.

1511

1512 Edge Cases:

- 1513 1. If the asked question has multiple answers and at least one of them aligns with
1514 the correct answer while none of them align with any of the other wrong answers,
1515 then provided that the top 5 conditions are met, we can mark the set as good.
- 1516 2. If two questions are very similar (even within different clips) and both are GOOD
1517 , only choose one as GOOD and reject the other one with a comment mentioning
1518 this. We do not expect this to happen more than 1 or 2 times in a 100.
- 1519 3. There might be more such edge cases, please feel free to contact me in such cases
1520 and we can expand.

1521 The teams were compensated at a rate of \$5/hour, a wage significantly higher than the minimum hourly
1522 rate in Bangladesh. This was done to ensure fair compensation for the complex tasks performed while
1523 also contributing to the quality of the work delivered. It's important to note that our collaboration
1524 with Quantigo followed ethical guidelines, with the fair treatment of all employees involved and the
1525 appropriate respect for their expertise and labor.

1526 **12 Benchmarking details**

1527 **12.1 Violet**

1528 Violet is a video language model comprised of a visual encoder, text encoder, and multimodal
1529 transformer pretrained on a variety of masked visual modeling tasks ranging from simple ones such as
1530 RGB pixel values up to more high levels ones such as spatially focussed image features. It performs
1531 competitively on a variety of video-language tasks such as Video-QA and Video-Text Retrieval. We
1532 evaluate one pre-trained model and 3 models finetuned on lsmdc-mc, msvtt-qa, and msvtt-retrieval.
1533 We evaluate using both 5 frames and 75 frames and choose the model with the best overall accuracy.

1534 **12.2 mPLUG-Owl**

1535 By default, mPLUG-Owl does not possess inherent capabilities for direct video question answering.
1536 As such, we undertook several experiments to adapt it to our required format. One approach involved
1537 inputting all answer choices in the form of a shuffled test. However, this resulted in a bias towards
1538 selecting the first option in most cases. For another approach, FrozenBiLM offered a methodology
1539 for frozen zero-shot models to operate in the context of multiple-choice video question answering,
1540 which inspired us to adapt this methodology for mPLUG-Owl. As mPLUG-Owl utilizes word-level
1541 tokenization, we could extract the confidence score for each generated token, particularly the 'Yes'
1542 token. We recorded the 'Yes' token confidence score for each answer option. In instances where the
1543 'Yes' token was absent, we assigned the confidence score as zero, though empirically, in most cases,
1544 the model output was positive and contained the 'Yes' token. Ultimately, we selected the answer
1545 option with the highest 'Yes' confidence score as the model output given the question. In scenarios
1546 where multiple options scored the same highest confidence for the 'Yes' token, we randomly selected
1547 the answer from these top-scoring options. It should be noted that mPlug-Owl was originally trained
1548 to process a single image, and its capacity to handle additional frames is an emergent ability that has
1549 not been thoroughly tested to date."

1550 **12.3 InternVideo**

1551 The two most closely aligned formats supported by InternVideo are open-ended Video Question
1552 Answering and Zero-shot Multiple Choice tasks. In the case of open-ended Video Question An-
1553 swering, the task is to predict the answer to a question posed within a video. However, due to
1554 the restricted vocabulary of open-ended answers in open-ended Video Questions Answering, we
1555 decided to formulate EgoSchema within the context of a Zero-shot Multiple Choice task. This task
1556 aims to identify the correct answer from a set of given options, without the inclusion of a question.

1557 InternVideo has provided finetuned weights for two datasets: MSRVTT and LSMDC. We selected
1558 the model finetuned on MSRVTT because it shares greater contextual similarity with EgoSchema.

1559 **12.4 Human**

1560 To conduct human benchmarking, we engaged the services of Quantigo, the same company responsi-
1561 ble for human data curation within our dataset. A distinct team of ten employees within the company
1562 was assigned to carry out human benchmarking on our dataset. The answers were randomized and
1563 presented to Quantigo in the form of a test. The following are the precise instructions provided to the
1564 annotators:

- 1565 - Setting 1: Unlimited setting -- The goal is to get answers as accurately as
1566 possible without worrying about time. For this setting do all the 100 clips. I
1567 will send the other CSVs and descriptions once this is done.
1568
- 1569 - Setting 2: 1 minute timed setting -- In this case, the test taker (annotator) has
1570 only 1 minute to spend per question (including watching video/reading text/
1571 everything). If they do not have the answer, just guess the best based on their
1572 intuition and move on. For this setting, please do all 100 clips.
1573
- 1574 - Setting 3: 3 minutes timed setting-- Same as above but with 3 min instead of just
1575 one. For this setting, please also do all 100 clips.
1576
- 1577 - Setting 4: Video -> Text setting -- In this case, the taker is not allowed to read
1578 the text before looking at the video and at the video after reading the text.
1579 In other words, the test taker can spend as much time as they want to look at
1580 the video first and then must move on to answering the question. They cannot go
1581 back to the video once they start reading the text. This is an untimed setting
1582 -- they can take as much time as they want per question. In this case, please
1583 limit the number of questions to 5 hours of human effort.
1584
- 1585 - Setting 5: 180 frames setting -- This is the same as the untimed setting except
1586 the annotator has access to only 1 frame per second (ie the video feels like a
1587 GIF with one frame per second instead of the usual 30 frames per second) --
1588 each video is still 3 minutes long, but it feels more jittery. All instructions
1589 remain the same as in an untimed setting. In this case also, please limit the
1590 number of questions to 5 hours of human effort.

1591 **13 Code of Ethics**

1592 **13.1 Research involving human subjects or participants**

1593 **13.1.1 Fair Wages**

1594 The Annotation teams were compensated at a rate of 5 dollars per hour, a wage significantly higher
1595 than the minimum hourly rate in Bangladesh. This was done to ensure fair compensation for the
1596 complex tasks performed while also contributing to the quality of the work delivered. It's important
1597 to note that our collaboration with Quantigo followed ethical guidelines, with the fair treatment of all
1598 employees involved and the appropriate respect for their expertise and labor.

1599 **13.2 Data-related concerns**

1600 **13.2.1 Privacy**

1601 For the visual part of our dataset, the Ego4D paper employed an array of deidentification procedures
1602 primarily centered on ensuring a controlled environment with informed consent from all participants,
1603 and, where applicable, in public spaces with faces and other personally identifiable information
1604 suitably obscured.

1605 **13.2.2 Consent**

1606 Ego4d paper privacy procedures have included obtaining informed consent from those wearing the
1607 cameras. Given that the video collection was conducted by Ego4D, we are not in a position to provide
1608 specific instructions that were given to the camera wearers. You can find the Ego4D privacy statement
1609 at <https://ego4d-data.org/pdfs/Ego4D-Privacy-and-ethics-consortium-statement.pdf>.

1610 **13.2.3 Deprecated datasets**

1611 We confirmed with the creators of Ego4D that the dataset will be maintained and accessible for the
1612 foreseeable future.

1613 **13.2.4 Copyright and Fair Use**

1614 EgoSchema will be publicly released under the Ego4D license, which allows direct public use of the
1615 video and text data for both research and commercial purposes.

1616 **13.2.5 Representative evaluation practice**

1617 Ego4D has strived to gather a diverse collection of videos, originating from 74 distinct locations
1618 across seven different countries. The dataset encompasses a broad variety of occupations, age groups,
1619 and genders. For instance, 96 of the participants are over the age of 50, and females constitute 45%
1620 of the participant population.

1621 **13.3 Societal Impact**

1622 **13.3.1 Safety**

1623 The safety and ethics of our dataset were fundamental to our research project from the outset. The
1624 dataset we created is a derivative of the Ego4D dataset, which was developed with stringent privacy
1625 and ethics standards. The Ego4D consortium required each partner to create and adhere to a policy that
1626 complied with their institutional research policy, including review by independent ethics committees
1627 where necessary. Additionally, they obtained informed consent from camera wearers, respected
1628 the rights of others in private spaces, and followed de-identification requirements for personally
1629 identifiable information (PII). With these stringent measures in place, we can confidently say that the
1630 potential for misuse of our dataset to harm, injure or kill people through its direct application, side
1631 effects, or potential misuse is significantly minimized.

1632 **13.3.2 Security**

1633 We thoroughly scanned the videos sourced from Ego4D and confirmed that they were free from any
1634 potential malware or viruses. In addition, to safeguard against any potential unauthorized access,
1635 we meticulously adhered to the security guidelines outlined by Github, which serves as our primary
1636 platform for website and data hosting.

1637 **13.3.3 Discrimination**

1638 Drawing from the rich diversity inherent in the Ego4D dataset, and subsequently carried over to
1639 our EgoSchema, we are confident that our dataset does not lend itself to discrimination based on
1640 factors such as race, ethnicity, age, occupation, or sex. However, we recognize that the potential
1641 for unforeseen biases always exists in any dataset. Our team is committed to maintaining an
1642 ongoing, vigilant monitoring process of the various applications of our dataset. Should we identify
1643 any indications of discriminatory usage or inherent bias stemming from our dataset, we pledge
1644 to swiftly implement corrective measures. Our aim is to ensure that our dataset remains a tool
1645 for inclusive, equitable, and bias-minimized research, contributing positively to the field of video
1646 question-answering systems.

1647 **13.3.4 Surveillance**

1648 Our data was collected with explicit consent from participants, and thoroughly de-identified to protect
1649 privacy and well-being. The dataset is intended for improving video question-answering systems, not
1650 for surveillance or predicting protected categories. We strongly advocate for its ethical use, consistent
1651 with all applicable laws and regulations, and discourage any misuse that could endanger individual
1652 well-being.

1653 **13.3.5 Deception and Harassment**

1654 Given the nature of the data and its intended use in question-answering tasks, it is unlikely that it
1655 could be exploited as a tool to promote hate speech or abuse. The dataset does not include content
1656 that could be construed as promoting hate speech. However, we recognize the potential for any
1657 technology to be misused. Therefore, we remain committed to monitoring the use of our dataset
1658 and taking appropriate action should we become aware of any misuse that contradicts our ethical
1659 guidelines.

1660 **13.3.6 Environment**

1661 The creation and use of our dataset primarily involve computational resources, which, while they do
1662 have an energy footprint, are a far cry from direct environmental exploitation or damage. Our dataset
1663 is designed with the intent to not necessitate high-end, energy-intensive GPUs for use. However, we
1664 acknowledge the broader concerns about the environmental impact of large-scale data processing
1665 and AI model training. Therefore, we recommend the use of energy-efficient hardware and practices
1666 where possible, such as optimizing algorithms for efficiency, utilizing cloud services that prioritize
1667 renewable energy, and considering the trade-off between computational resource usage and model
1668 performance.

1669 **13.3.7 Human Rights**

1670 Our dataset is firmly rooted in respect for human rights and does not facilitate any form of illegal
1671 activity. Participant privacy was ensured through robust de-identification procedures. We condemn
1672 any misuse of our dataset that infringes upon rights to privacy, speech, health, liberty, security, legal
1673 personhood, or freedom of conscience or religion. We urge researchers to use our dataset ethically,
1674 legally, and in alignment with its intended purpose of furthering video question-answering research.

1675 **13.3.8 Bias and fairness**

1676 Our dataset, sourced from Ego4D, may carry forward its inherent biases. Additional biases could
1677 be introduced through text data from Language Models and human curation. We've taken steps to
1678 minimize errors but recognize that some biases may persist. These could influence model performance
1679 and potentially perpetuate bias against certain groups. We plan to mitigate these risks by hosting a
1680 crowd-sourced errata board, allowing the open-source research community to help continually correct
1681 and refine our dataset.

1682 **13.4 Impact Mitigation Measures**

1683 **13.4.1 Data and model documentation**

1684 Dataset documentation is provided in the EgoSchema datasheet. Additional details will be provided
1685 on our [project page](#) where users will be able to explore and download the dataset directly.

1686 **13.4.2 Data and model licenses**

1687 EgoSchema will be publicly released under the Ego4D license, which allows direct public use of the
1688 video and text data for both research and commercial purposes.

1689 **13.4.3 Secure and privacy-preserving data storage and distribution**

1690 the Ego4D paper employed an array of deidentification procedures primarily centered on ensuring a
1691 controlled environment with informed consent from all participants, and, where applicable, in public
1692 spaces with faces and other personally identifiable information suitably obscured.

1693 **13.4.4 Allowing access to research artifacts**

1694 The code used for generation will be provided and full instructions on how to run the code our GitHub
1695 page: <https://github.com/egoschema/EgoSchema>

1696 **13.4.5 Disclose essential elements for reproducibility**

1697 The code used for benchmarking will be made available with added instructions on how to reproduce
1698 the results on our GitHub page: <https://github.com/egoschema/EgoSchema>

1699 **13.4.6 Ensure legal compliance**

1700 The Ego4D consortium collected the dataset locally ensuring legal compliance for each region. We
1701 also publish EgoSchema under the Ego4D license and hence ensure legal compliance by extension.

